Département de géomatique appliquée
Faculté des lettres et sciences humaines
Université de Sherbrooke

Comparaison des stratégies directe et indirecte pour la cartographie du volume forestier : cas de la forêt boréale à Terre-Neuve, Canada.

Mélodie Bujold

Présentation du mémoire
pour l'obtention du grade de Maître ès sciences géographiques (M.Sc.),
cheminement recherche en environnement, Géomatique.

Août 2019

Avant-propos

Ce projet de maîtrise s'intègre dans un projet plus vaste, collaboratif entre les Universités, le gouvernement et l'industrie, nommé AWARE (Assesement of Wood Attributes from Remote Sensing), qui vise à utiliser la télédétection afin d'améliorer l'inventaire forestier du Canada et d'améliorer la modélisation des écosystèmes forestiers. Les trois thèmes abordés par AWARE sont les suivants : 1 : Les principaux facteurs de la structure des peuplements forestiers dans les régions forestières du Canada; 2 : Estimation de la variation structurelle, du volume et des espèces au niveau du peuplement et des arbres dans les sites cibles régionaux / de gestion; et 3 : Dérivation des caractéristiques des arbres individuels dans les parcelles et de leurs liens avec les attributs de la fibre de bois, pour des inventaires forestiers améliorés. Le présent projet visait à répondre à la question 3 du thème 1 d'AWARE : « How do we effectively derive sample based estimates along LiDAR transects and then scale-up, and model error propagation of these estimates, to produce information products useful for regional and national level reporting? ». Le projet a été réalisé en collaboration avec le Service canadien des forêts (SCF), division de l'Atlantique (une division de Ressources naturelles Canada), l'Université de Sherbrooke, le Forestry et Agrifood Agency de Terre-Neuve et la Kruger (Corner Brook Pulp and Paper).

Identification du jury

Directeur de recherche : Prof. Richard Fournier
Centre d'Applications et de Recherches en Télédétection (CARTEL), Département de géomatique appliquée, Université de Sherbrooke, Sherbrooke, QC, Canada.

Codirectrice de recherche : Joan Luther
Ressources Naturelles Canada, Centre de Foresterie de l'Atlantique, Corner Brook, T.-N.-L., Canada.

Membres du jury :

Jury interne : Prof. Kalifa Goïta
Centre d'Applications et de Recherches en Télédétection (CARTEL), Département de géomatique appliquée, Université de Sherbrooke, Sherbrooke, QC, Canada.

Jury externe : Luc Guindon
Ressources Naturelles Canada, Centre de Foresterie des Laurentides, Sainte-Foy, QC, Canada.

Résumé du projet

La gestion des ressources forestières sur de grands territoires requiert une cartographie précise des attributs de la forêt. Les programmes d'inventaire forestiers traditionnels s'appuient sur des photographies aériennes et des interprètes expérimentés pour cartographier les attributs forestiers. Cela nécessite une grande quantité de ressources. Une stratégie directe de cartographie alternative consiste à établir des relations statistiques entre (i) les attributs mesurés dans des placettes forestières et (ii) les données spectrales des images satellitaires combinées à des couches spatiales disponibles (par ex., relief, climat). Malheureusement, l'utilisation de l'imagerie multispectrale à elle seule n'atteint souvent pas un niveau de précision satisfaisant à cause de la saturation connue du signal optique à fort volume forestier. La donnée LiDAR aéroportée (ALS) sur une portion du territoire peut être utilisée pour améliorer la précision des cartes forestières par une stratégie de cartographie indirecte à deux phases. La phase 1 combine les mesures des placettes et la donnée ALS. La phase 2 combine la carte produite à la phase 1 avec des données satellitaires/spatiales pour une carte du territoire étendu. Ce projet vise à comparer la prédiction du volume total de bois de l'île de Terre-Neuve obtenue selon les stratégies directe et indirecte, ainsi qu'à comparer les approches de modélisation statistiques paramétriques et non-paramétriques (regression Ordinary least squares (OLS) vs Random Forest (RF)) pour chacune des stratégies.

Les modèles de la stratégie indirecte utilisés pour prédire le volume total sur les placettes de validation, basées sur les données ALS, expliquaient systématiquement une faible variance (16 % pour OLS et 11 % pour RF), avec des erreurs de prédiction relatives élevées pour les deux approches (47 % et 50 %). Les modèles de la stratégie directe, développés avec les placettes au sol, expliquaient une variance similaire de celle obtenue par la stratégie indirecte pour les deux approches (11 % et 14 %), avec des erreurs de prédiction tout aussi élevées (50 % et 56 %). Les modèles développés selon la stratégie indirecte n'ont pas entrainé une augmentation significative de la correspondance entre les valeurs observées et prédites, et ce, pour les deux approches. Ces résultats peuvent être expliqués par de nombreux facteurs, tels que le faible niveau de représentativité des placettes, la résolution différente des images satellitaires ou la densité de points des données ALS.

Mots clés : Modélisation spatiale, LiDAR, Inventaire forestier, Volume de bois, Imagerie, Forêt boréale, Régression, Random Forest.

Remerciements

J'aimerais tout d'abord profiter de l'occasion pour remercier le professeur Richard Fournier, de l'Université de Sherbrooke, et ma co-directrice Joan Luther, du Service Canadien des Forêts de Terre-Neuve, de m'avoir proposé ce projet et pour tout le soutien qu'ils m'ont toujours apporté durant la réalisation de cette maîtrise. J'aimerais également remercier Olivier Van Lier de m'avoir fourni les multiples données nécessaires à mon projet et d'avoir toujours été disponible pour répondre à mes questions. Je tiens aussi à remercier Mike Wulder et Joanne White du Service canadien des forêts de m'avoir fourni les images composites du « Best available pixel ». Ces images composites ont été générées pour le projet « National Terrestrial Ecosystem Monitoring System (NTEMS) » : Surveillance intersectorielle nationale opportune et détaillée, cofinancées par le Programme d'initiatives liées au gouvernement de l'Agence spatiale canadienne et le Service canadien des forêts de Ressources naturelles Canada. Un grand merci également à Doug Bolton, de l'Université UBC, pour son grand support dans la recherche de solutions pour l'amélioration des résultats de modélisation préliminaires. Je tiens aussi à remercier Marc Mazerolle et François Rousseu pour leur appui dans le volet statistique des méthodes d'échantillonnage. Enfin, le projet n'aurait pas été concrétisé sans l'apport financier d'AWARE, le support de Barry Elkins comme partenaire privé de la Kruger (Corner Brook Pulp & Paper) et de Boyd Pittman and Scott Payne de NL Forestry & Agrifoods Agency.

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

CCMEAN: Percentage of all returns above mean

CreliefRATIO: Canopy relief ratio

LH99: 99th percentile> 2m

LHCURTmeanCUBE: Elevation CURT mean CUBE

LHKURT: Elevation kurtosis

LHL2: Elevation L2-moment

LHL4: Elevation L4-moment

LHLCOV: Second L-moment ratio (coefficient of variation) of point heights> 2 m

LHLKURT: Fourth L-moment ratio (coefficient of kurtosis) of point heights> 2 m

LHMADMODE: Elevation median of the absolute deviations from the overall mode (MAD)

LHMIN: Elevation minimum

LJLSKEW: Elevation L-moment skewness

OLS: Ordinary least squares

RF: Random Forest

RATIOMEAN: Number of returns above mean height/Total first returns *100

RATIOMODE. Number of returns above mode height/Total first returns *100

1. Introduction

La planification, la gestion et la surveillance des ressources forestières de grands territoires forestiers, dans un contexte de gestion durable des ressources, requièrent des cartes précises des attributs forestiers, à savoir la composition et la structure des peuplements (McRoberts et al., 2010). Ces informations proviennent encore souvent des placettes de terrain collectées par les inventaires forestiers conventionnels (Tomppo et al., 2010). Cependant, il existe d'importantes limites concernant la collecte parcellaire, telles que leur coût, le manque de couverture spatiale et la longueur du cycle des mises à jour. Pour faire face à ces problèmes de collecte de données, les praticiens ont souvent eu recours à des experts de l'interprétation de photographies aériennes, supportée par les placettes terrain, afin d'extraire les attributs du peuplement forestier, y compris la composition des espèces, les classes de hauteur, la fermeture du couvert forestier (crown closure), l'âge et la qualité du site (Avery and Burkhart, 2015). Pour les zones forestières inaccessibles et éloignées, l'acquisition de mesures précises ayant trait à la structure de la végétation dans un cadre opérationnel est encore plus problématique et demeure donc un défi constant pour les gestionnaires.

Ces dernières décennies, il y a eu un grand intérêt pour le développement de méthodes reposant sur l'imagerie optique afin d'extrapoler les données forestières structurelles au-delà de la couverture des données ALS ou de terrain afin de représenter une région entière. Les progrès réalisés dans l'utilisation des données de télédétection offrent des possibilités de cartographie détaillée et précise des caractéristiques de la forêt à un coût inférieur à celui des pratiques actuelles d'inventaire conventionnel (Tomppo et al., 2008). Les images optiques satellitaires peuvent être combinées avec les valeurs des placettes de terrain pour prédire les attributs forestiers de chaque pixel sur de grandes superficies (Tuominen et al., 2003 ; Tomppo et al., 2008 ; Wulder et al., 2008 ; Leboeuf et Fournier, 2015). L'étude de Tomppo et al. (2008) stipule que l'erreur relative (RMSE %) des prédictions du volume effectuées au niveau du pixel par les recherches en Suisse et en Norvège demeure cependant élevée, soit entre 50-80 %. De plus, de nombreuses études ont démontré que l'utilisation de données de LiDAR aéroporté (Airborne Laser Scanning : ALS) peut améliorer la prédiction de plusieurs attributs forestiers, dont la hauteur moyenne et dominante, le diamètre moyen des tiges à hauteur de poitrine, la surface terrière, le nombre de tiges, le volume et la biomasse (Hudak et al., 2002 ; Packalén et Maltamo, 2006 ; Holmgren et al., 2008 ; McInerney et al., 2010 ; Bouvier et al., 2015). L'amélioration des modèles prédictifs à l'aide des données ALS est un domaine de recherche actif.

Les stratégies cartographiques des attributs forestiers sur de vastes territoires utilisant de la télédétection peuvent être divisées en deux catégories. Premièrement, la stratégie la plus utilisée jusqu'à maintenant est une stratégie directe qui consiste à développer des relations empiriques entre les mesures des placettes terrain et des données de télédétection ayant une couverture complète, telles que les données ALS, l'imagerie satellitaire et d'autres données spatiales auxiliaires (Næsset, 2002; Hudak et al., 2006; McRoberts, 2006; McInerney et al., 2010; Wulder et al., 2012). Le succès de la stratégie directe est donc étroitement lié à la disponibilité d'un ensemble étendu de placettes de terrain dans la zone d'intérêt. Les relations empiriques sont utilisées pour prédire et cartographier les attributs forestiers du territoire. Des stratégies directes ont été adoptées ou envisagées par plusieurs pays pour leurs inventaires forestiers nationaux : Finlande (Tomppo, 1991), Norvège (Nilsson, 1997), Nouvelle-Zélande (Tomppo et al., 1999), États-Unis (Franco-Lopez et al., 2001 ; McRoberts et al., 2002), la Chine (Tomppo et al., 2001) et certaines provinces du Canada (Beaudoin et al., 2014, Leboeuf et Fournier, 2015). Par exemple, le chercheur norvégien Naesset (2002) et le groupe de chercheur Hudak et al. (2006) ont utilisé du LiDAR mur-à-mur (c'est-à-dire sur l'ensemble du territoire concerné) dans le cadre d'une stratégie directe afin d'améliorer la précision des inventaires forestiers. Les modèles de régression de Naesset, pour la prédiction de 6 attributs forestiers d'une forêt de conifères, ont expliqué entre 60-97 % de la variabilité des valeurs de référence au sol des six caractéristiques étudiées. Le biais et la déviation standard des différences entre les valeurs prédites et les valeurs des parcelles de référence (entre parenthèses) variaient selon les attributs : -0,58 à -0,85 m (0,64-1,01 m) pour la hauteur moyenne, -0,60 à -0,99 m (0,067-0,84 m) pour la hauteur dominante, 0,15-0,74 cm (1,33-242 cm) pour le diamètre moyen, 34-108 ha^{-1} (97 — 466 ha^{-1}) pour le nombre de tiges, 0,43-2,51 $m^2 ha^{-1}$ (1,83-3,94 $m^2 ha^{-1}$) pour la surface terrière et enfin, 5,9-16,1 $m^3 ha^{-1}$ (15,1-35,1 $m^3 ha^{-1}$) pour l'attribut du volume. Ces résultats démontrent que les données LiDARs mur-à-mur utilisées selon une stratégie directe peuvent contribuer à obtenir des prédictions précises. Cependant, l'acquisition de données ALS sur une zone entière n'est pas toujours justifiable ou possible en raison des coûts élevés, en particulier pour de grands territoires.

Il existe un besoin pour une stratégie alternative, appelée indirecte, qui tire parti des ensembles de données à deux niveaux : la donnée ALS sur une fraction du territoire et la donnée satellitaire optique sur l'ensemble du territoire. En effet, la prédiction des attributs forestiers sur de grands territoires peut utiliser les données ALS comme un outil d'échantillonnage dans une approche en deux étapes. Cette stratégie peut être considérée comme une stratégie indirecte dans laquelle deux ensembles de

relations empiriques sont développés (Andersen et al., 2012 ; Strunk et al., 2014) : (1) le premier ensemble de relations est construit entre les placettes disponibles sur le terrain et les données ALS acquises seulement sur une portion du territoire, et (2) le second ensemble de relations est construit entre les valeurs échantillonnées provenant de la prédiction sur les zones couvertes par les données ALS et les couches spatiales couvrant entièrement la zone d'intérêt, y compris les images satellitaires optiques, les données topographiques ou climatiques. La prédiction des attributs forestiers est ensuite généralisée à partir de la zone couverte par les données ALS aux autres zones incluses dans la zone d'intérêt. Récemment, l'intérêt pour la stratégie indirecte s'est accru parce que la possibilité de recueillir des données à diverses résolutions et échelles offre l'opportunité d'améliorer la prédiction des attributs forestiers pour les grandes surfaces, sans le coût excessif d'acquisition des données ALS mur-à-mur. Le coût d'acquisition des données supplémentaires peut être compensé si la précision de la prédiction obtenue par la stratégie indirecte est supérieure à celle de la stratégie directe.

Une des premières études ayant exploré l'intégration de transects ALS dans le but de cartographier les emplacements non couverts par l'ALS a été réalisée en Oregon en 2002 par Hudak et al.. Ces auteurs ont testé cinq méthodes aspatiales et spatiales pour étendre la prédiction de la hauteur de la canopée et ont conclu que l'intégration des données ALS et Landsat améliorait l'utilité des deux ensembles de données. En effet, les R^2 obtenues sont majoritairement élevés selon la technique d'échantillonnage testée, soit entre 0,53 et 0,94. D'autres études ont utilisé les estimés dérivés des données ALS comme substituts de parcelles terrain et ont étendu ces estimations avec l'utilisation d'imagerie satellitaire. En autres, McInerney et al. (2010) et Pascual et al. (2010) ont extrapolé les hauteurs de canopée dérivées des prédictions effectuées sur les données ALS en utilisant soit de l'imagerie de résolution moyenne provenant du satellite indien de télédétection ou les données des bandes et les indices spectraux du satellite « Landsat Enhanced Thematic Mapper (ETM +) ». Maselli et al. (2011) a quant à lui évalué à la fois la régression et k-NN pour étendre l'estimation du volume de tiges dérivées de la hauteur moyenne des peuplements obtenus de transects ALS sur des images Landsat ETM+. Par la suite, les concepts de « placettes LiDAR » (Wulder et al., 2012 ; Zald et al., 2016) et « échantillonnage LiDAR » (examinés par Wulder et al., 2012) ont été proposés pour atténuer le besoin de données de parcelles terrain pour la caractérisation et la cartographie des forêts sur de grands territoires.

La base conceptuelle d'une stratégie indirecte (à plusieurs niveaux) combinant des placettes au sol, des données ALS et des données satellitaires pour l'inventaire forestier sont décrites en détail par

Andersen et al. (2012) et aussi démontré par d'autres chercheurs (Strunk et al., 2014 ; Holm et al., 2017). Andersen et al. (2012), ont estimé la biomasse totale pour les forêts boréales de l'intérieur de l'Alaska en implémentant une stratégie multiniveau (indirecte) qui combine des parcelles terrains, un échantillonnage au sein des transects ALS et de l'imagerie satellitaire. Strunk et al. (2014) ont aussi conclu que la stratégie indirecte, impliquant des données LiDAR seulement sur une portion du territoire, améliore l'estimation des attributs forestiers (biomasse, surface terrière et le nombre d'arbres) par rapport aux estimations obtenues par l'emploi d'une stratégie directe utilisant seulement les données issues de l'imagerie Landsat. En effet, ils ont démontré que l'approche indirecte (approche multiniveau) a entrainé une réduction de la variabilité résiduelle jusqu'à 36 % pour l'attribut de la biomasse. Enfin, d'autres ont combiné le LiDAR de terrain, l'aérien et le LiDAR spatial (Système Laser Altimètre géoscientifique [GLAS]) pour produire des cartes régionales des attributs de peuplement (Mahoney et al., 2018) et pour estimer la biomasse au-dessus du sol et le carbone des forêts boréales à l'aide de stratégies d'échantillonnage à plusieurs niveaux (Boudreau et al., 2008 ; Holm et al., 2017). Bien que ces études aient démontré une amélioration de la précision de la cartographie et de l'estimation de certains attributs forestiers en utilisant une stratégie indirecte (plusieurs niveaux) de modélisation, il est nécessaire de tester davantage les capacités comparatives des deux stratégies avec différents jeux de données, à des échelles différentes et dans des environnements forestiers différents.

La cartographie des attributs forestiers sur de grands territoires, et d'une forêt boréale, implique de nombreux défis. D'abord, les variations de structure existant dans une forêt peuvent induire des difficultés à prédire avec précision les attributs de la forêt avec un seul modèle. La mise en place de modèles multiples par type d'écozone peut s'avérer nécessaire (Matasci et al., 2018). L'usage de la latitude et de la longitude comme variables explicatives du modèle peut aussi être nécessaire pour informer le modèle sur les conditions écologiques localisées (Matasci et al., 2018). La représentativité spatiale des données LiDARs est un autre critère important pour la génération de cartes précises des attributs de la forêt sur de grands territoires. En effet, la disponibilité de placettes dérivées du LiDAR est un facteur clé pour le développement du modèle. Une autre condition essentielle pour la cartographie sur une grande surface est de posséder des données optiques cohérentes et spatialement complètes sur le territoire. Des indices spectraux dérivés des pixels de l'imagerie satellite sont cruciaux pour la cartographie. Ainsi, la représentativité de la réflectance de surface de l'image est essentielle pour garantir que les peuplements présentant des caractéristiques

similaires dans la zone cartographiée présentent également des valeurs spectrales associées similaires. Afin de combler les lacunes spectrales et temporelles de l'image et d'ainsi obtenir une couverture complète nécessaire à la modélisation, le produit de Landsat appelé « Best avalable pixel » (BAP) utilise des fonctions de notation des pixels, décrites en détail dans White et al. (2014), pour créer une image composite. Ces fonctions attribuent une note à chaque pixel pour (i) le capteur, (ii) le jour d'acquisition de l'année, (iii) la distance aux nuages et l'ombre des nuages, (iv) l'opacité atmosphérique. La valeur de réflectance de surface du pixel ayant la somme la plus élevée des quatre scores est alors utilisée dans l'image composite BAP pour une année donnée et les pixels sans observations appropriées sont assignés à des valeurs de « No Data » et sont adressés plus tard dans le processus de composition de l'image.

Les approches basées par pixels utilisant des données de télédétection (ALS ou imagerie satellite) impliquent le développement de modèles pour prédire les attributs forestiers d'intérêt pour chaque pixel de la région visée. Conséquemment, l'amélioration de la précision de prédiction des attributs forestiers des stratégies directes et indirectes peut dépendre de l'approche de modélisation utilisée. Les méthodes utilisées pour la prédiction et la cartographie des attributs forestiers à partir de données géospatiales sont nombreuses et variées. Les approches de modélisation paramétrique et non-paramétrique sont couramment utilisées pour prédire les attributs forestiers à partir de l'imagerie satellitaire, des données ALS et des placettes terrain, à partir de stratégies directes ou indirectes. De nombreuses approches paramétriques sont courantes dans la littérature forestière, y compris l'estimation par maximum de vraisemblance (Hagner et Reese, 2007 ; Baatuuwie et Van Leeuwen, 2011), l'analyse discriminante (Thenkabail et al., 2004 ; Van Aardt et al., 2008) et la régression (le Maire et al., 2011 ; Tonolli et al., 2011). La régression est l'une des méthodes paramétriques les plus courantes pour lesquelles une hypothèse clé est que le modèle décrit correctement la population d'intérêt (Penner et al., 2013). En foresterie, les équations de régression linéaire multiple résolues à l'aide de la méthode des moindres carrés ordinaires (ordinary least squares: OLS) constituent l'une des méthodes paramétriques les plus couramment utilisées (par ex., Means et al., 2000 ; Hudak et al., 2006 ; McRoberts, 2006 ; Næsset, 2007 ; Andersen et al., 2012 ; Strunk et al., 2014). Cependant, la nature de plusieurs attributs forestiers complique l'emploi de cette méthode. Les prémisses de base de la régression OLS à propos des erreurs ne sont souvent pas respectées et des mesures pour y remédier doivent être prises, comme la transformation de la variable réponse (Frazer et al., 2011). De plus, la régression OLS suppose qu'il n'y a pas d'erreur de mesure

dans les variables explicatives, ce qui est souvent peu réaliste, surtout en ce qui concerne les variables de télédétection (Curran et Hay, 1986 ; Berterretche et al., 2005).

Une méthode non-paramétrique couramment utilisée pour la cartographie des attributs forestiers est celle du plus proche voisin (Nearest Neighbor : NN ou multiple NN : k-NN) (Katila et Tomppo, 2001 ; Maltamo et al., 2006 ; Packalén et Maltamo, 2007 ; Hudak et al., 2008 ; Tomppo et al., 2008 ; Falkowski et al., 2010 ; Maselli et al., 2011 ; Andersen et al., 2012 ; Beaudoin et al., 2014). Les techniques NN prédisent la variable réponse dans une unité cible non échantillonnée (par exemple, un pixel) en calculant une distance métrique statistique entre la cible et les échantillons de référence, ou « voisins », puis attribuent la valeur du voisin le plus proche à l'unité cible. De multiples métriques pour calculer la distance peuvent être utilisées pour prédire les attributs forestiers. En effet, à part la distance Euclidienne et Mahalanobis, la distance du voisin le plus similaire (Most similar neighbor: MSN), le gradient du plus proche voisin (The gradient nearest neighbor: GNN) ou l'incorporation de Random Forest (RF) au K-NN (RF-k-NN) peuvent entre autres être employés avec la méthode NN (Brosofske et al., 2013). Une des raisons de la popularité de cette méthode est sans contredit sa facilité à être comprise et appliquée. Cependant, le k-NN peut être exigeant en temps de calcul, en particulier avec des ensembles de données très volumineux, lorsque les attributs de la forêt sont cartographiés à des résolutions fines sur de grands territoires géographiques. De plus, cette méthode suppose que les placettes sont bien représentées et réparties sur l'ensemble du territoire couvert par les données de télédétection pour obtenir de bons résultats (Labrecque et al., 2006). Une autre méthode non-paramétrique, Random Forest (RF), gagne en popularité (Liaw and Wiener, 2002; Penner et al., 2013). La méthode RF consiste en une méthode d'ensemble qui combine les prédictions d'une myriade d'arbres de décisions individuelles, chacun basé sur une valeur de seuil appliquée à une variable prédictive (Breiman, 2001). Un avantage considérable de la méthode RF est sa capacité à gérer de grands ensembles de données et de nombreuses variables ainsi que d'être robuste au surajustement (Liaw et Wiener, 2002 ; Baccini et al., 2004 ; Prasad, 2006 ; Houghton et al., 2007).

La littérature fait aussi référence à l'utilisation d'autres techniques pour prédire les attributs forestiers. Dans les approches non-paramétriques, il est fait mention du parent de RF, soit la construction d'arbres de classification et de régression (CARTs), et la régression multivariée par spline adaptative (MARS). Enfin, l'Artificial Neural Networks (ANNs) (Fitzgerald et Lees, 1993 ; Rogan et al., 2008, pour la classification, Ingram et al., 2005 ; Niska et al., 2010 pour la prédiction de variables continues), la classification de l'occupation du territoire (Land Cover Classification: LCC)

(Labreque et al., 2006), l'étiquetage de cluster à l'aide de Structure et type pour la biomasse (BioCLUST) (Luther et al., 2006) sont d'autres méthodes qui ont fait l'objet d'étude pour la prédiction d'attributs forestiers.

Enfin, selon la littérature, l'approche non-paramétrique devient plus courante pour la cartographie forestière que l'approche paramétrique en raison de sa facilité de mise en œuvre, de l'absence d'hypothèses restrictives et de la capacité de certains algorithmes non-paramétriques à inclure des variables prédictives et réponses catégoriques (Evans and Cushman, 2009; Stumpf and Kerle, 2011). Seules quelques études comparent les approches paramétriques et non-paramétriques pour prédire et cartographier les attributs forestiers, par conséquent, on ne sait pas dans quelle condition chacune devrait être utilisée (Chirici et al., 2008 ; Andersen et al., 2012 ; Penner et al., 2013). Les résultats des études comparatives à ce jour ne sont pas unanimes quant aux améliorations possibles que peut apporter une approche non-paramétrique pour la prédiction des attributs forestiers. En effet, l'étude d'Andersen et al. (2012) portant sur l'estimation de la biomasse de la forêt boréale en Alaska a démontré que l'utilisation d'une approche de modélisation non-paramétrique (imputation par NN) réduisait l'erreur type de 7,3 % à 5,1 %. De l'autre côté, Chirici et al. (2008) ont quant à eux obtenu une diminution de l'erreur (RMSE) de 6 % et 7 % lors de la prédiction du volume de croissance d'une forêt alpine et méditerranéenne en Italie selon une approche paramétrique (régression) par rapport aux erreurs obtenues par la méthode k-NN. L'étude de Penner et al. (2013) en l'Ontario a obtenu des résultats similaires pour les méthodes de modélisation par régression et par RF lors de l'estimation de la surface terrière (RMSE de 1,14 pour la régression et de 0,96 pour RF). En somme, la détermination de la meilleure technique de modélisation dépend de plusieurs facteurs qui peuvent être subjectifs, y compris le but de l'analyse, les ressources disponibles, les caractéristiques des données, les conditions biophysiques sous-jacentes de la forêt et le type de variables réponses (Powell et al., 2010). La variabilité des résultats obtenus au sein des études comparatives des approches de modélisation démontre que d'autres études sont nécessaires.

2. Objectifs de recherche et hypothèses

L'objectif principal de ce projet est de développer une méthode efficace de cartographie pour l'attribut du volume total (Tvol) forestier pour la forêt boréale sur l'île de Terre-Neuve, au Canada. Les objectifs spécifiques sont : (i) comparer les stratégies de modélisation directe et indirecte qui combinent des placettes forestières, des transects ALS et des données à couverture complète pour cartographier l'attribut du Tvol, (ii) comparer les performances des méthodes paramétriques (OLS) et non-paramétriques (RF) pour prédire et cartographier l'attribut du Tvol sur un grand territoire.

Notre première hypothèse est qu'une stratégie indirecte, utilisant des transects de données ALS, améliorera la prédiction du Tvol par rapport à une stratégie directe, qui n'utilise que des placettes terrain et des couches de données à couverture complète (Andersen et al., 2012). Notre deuxième hypothèse est que la méthode RF augmentera la précision de la prédiction du Tvol de la forêt par rapport à la méthode OLS pour les stratégies de modélisation directe et indirecte (Penner et al., 2013).

3. Matériels

3.1 Zone d'étude

L'étude est réalisée sur l'île de Terre-Neuve, au Canada, centrée autour de 48° 32' 30" N et 56 ° 07' 30" W, et couvre 111 390 km^2, soit 11 millions d'hectares (Figure 1). Les forêts de Terre-Neuve sont situées à la partie la plus à l'est de la région de la forêt boréale de l'Amérique du Nord (Rowe, 1972). La végétation de conifères prédomine sur l'île, avec le sapin baumier (Abies balsamea (L.) Mill) et l'épinette noire (Picea mariana (Mill.) Britton, Sterns & Poggenb.) comme espèces d'arbres prédominantes. Les régions de l'ouest, du nord et de l'est de Terre-Neuve sont dominées par des peuplements de sapin baumier à cause de leur préférence pour les sols humides et bien drainés. Les peuplements d'épinettes noires ont tendance à dominer la région centrale, où les feux de forêt sont fréquents et ils forment environ un tiers de la forêt de l'île. La topographie de l'île de Terre-Neuve varie entre des vallées plates et des terrains relativement accidentés. La région intérieure sud des terres consiste en un plateau vallonné élevé. L'altitude la plus élevée se situe sur la côte ouest, atteignant 814 mètres au-dessus du niveau de la mer. La côte est se caractérise quant à elle par une topographie irrégulière, rugueuse et stérile, avec des pics de plus de 250 m de haut (Newfoundland government, 2015). Terre-Neuve est influencée par un climat maritime. Les températures moyennes annuelles sur l'île sont entre -3 à 10 °C et les plages de précipitations annuelles sont de 952 à 1686 mm (Résultats actuels de Nexus, 2015). Près de 75 % des précipitations annuelles tombent en pluie et le reste sous forme de neige.

3.2 Jeux de données

3.2.1 Placette terrain

Les sources de données des placettes terrain utilisées dans cette étude ont déjà été décrites par Luther et al. (2013). Un total de 62 placettes d'échantillonnages permanentes (Permanent Sample Plot : PSP) dominées par le sapin baumier et l'épinette noire se trouvaient dans la zone d'étude et ont été recoupées par l'étendue de l'acquisition des données ALS (Figure 1et Figure 2).

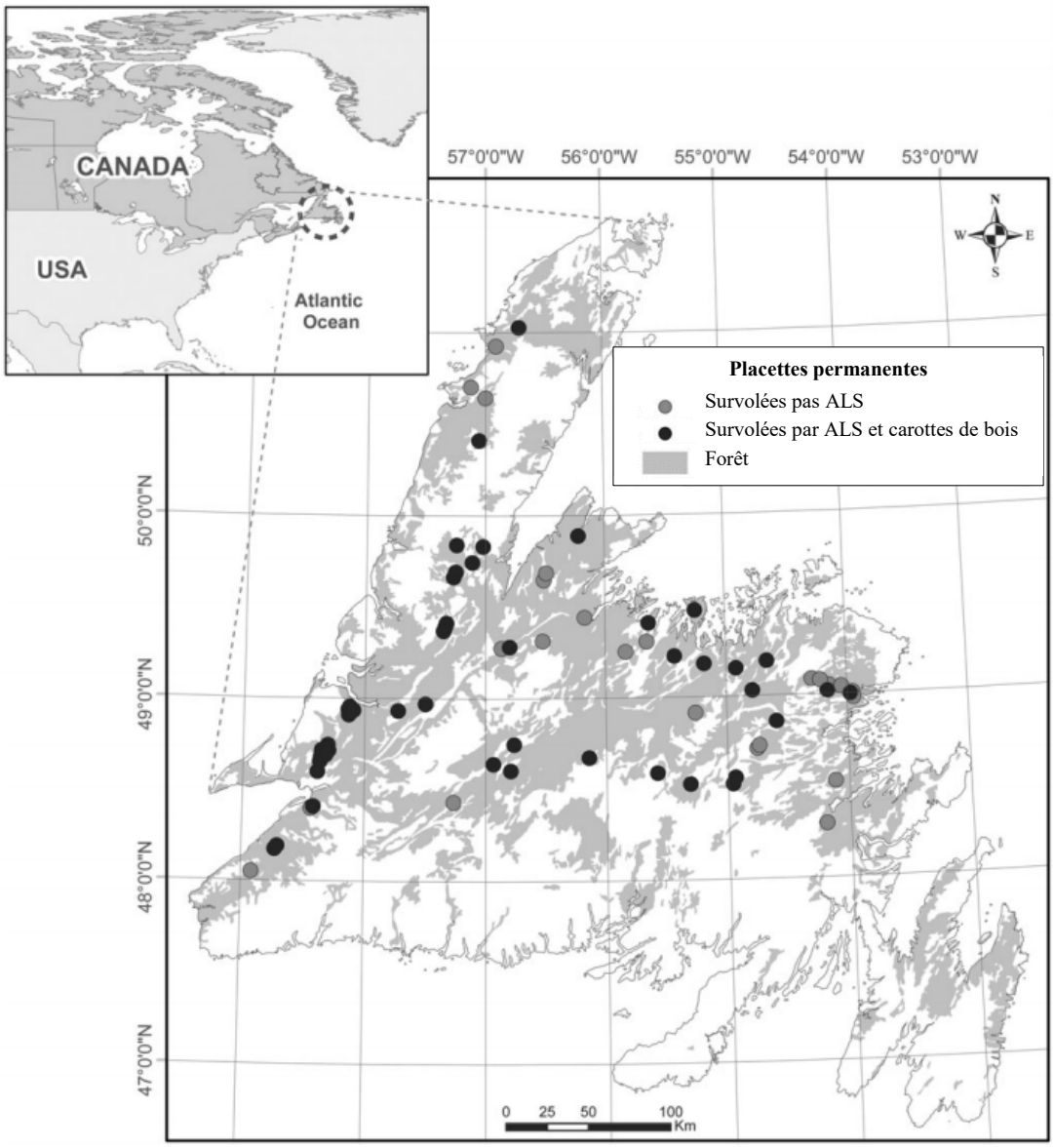


Figure 1. Étendue spatiale de la zone géographique indiquant les zones de forêt et les emplacements des placettes terrain (Luther et al., 2013).

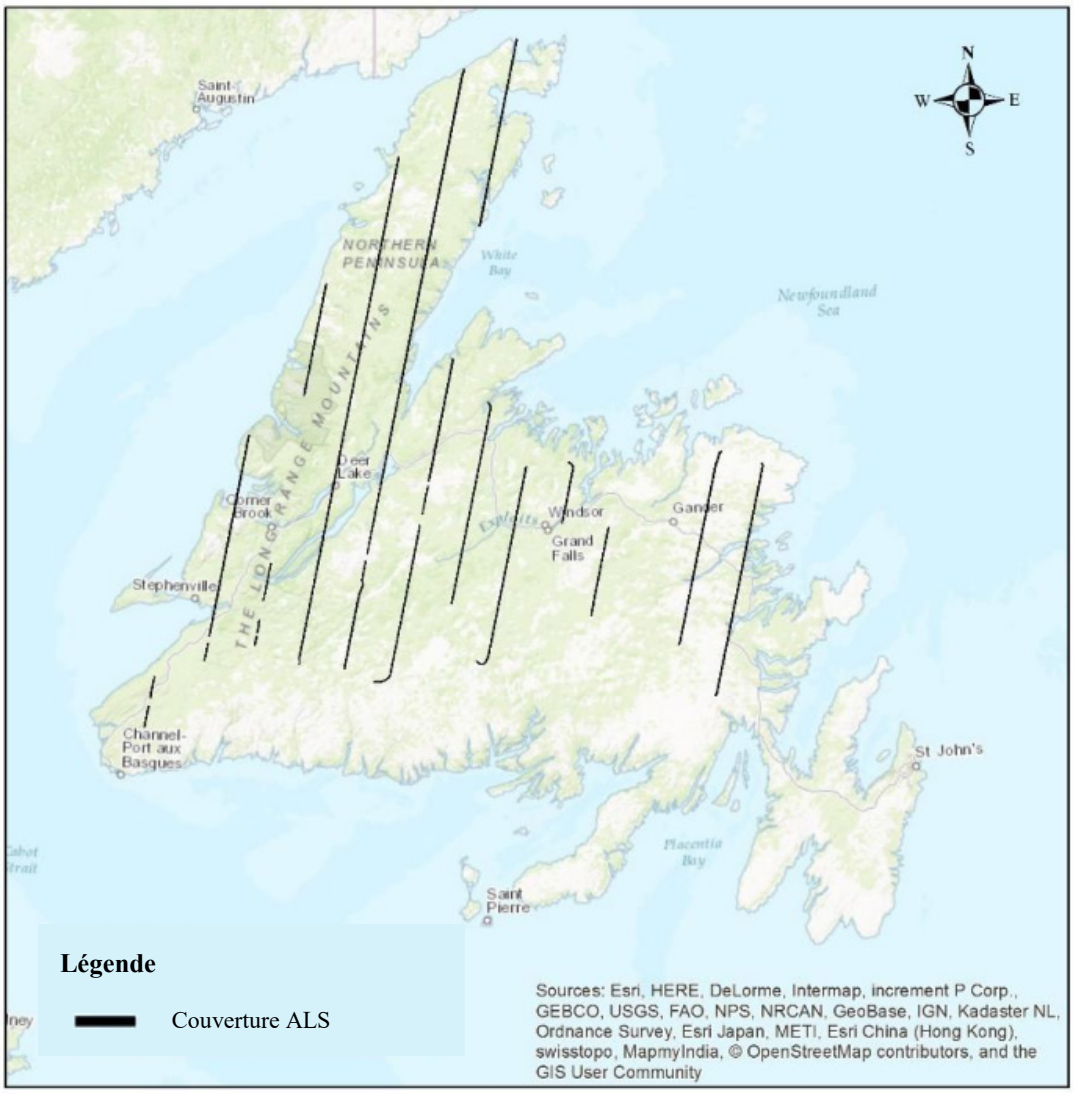
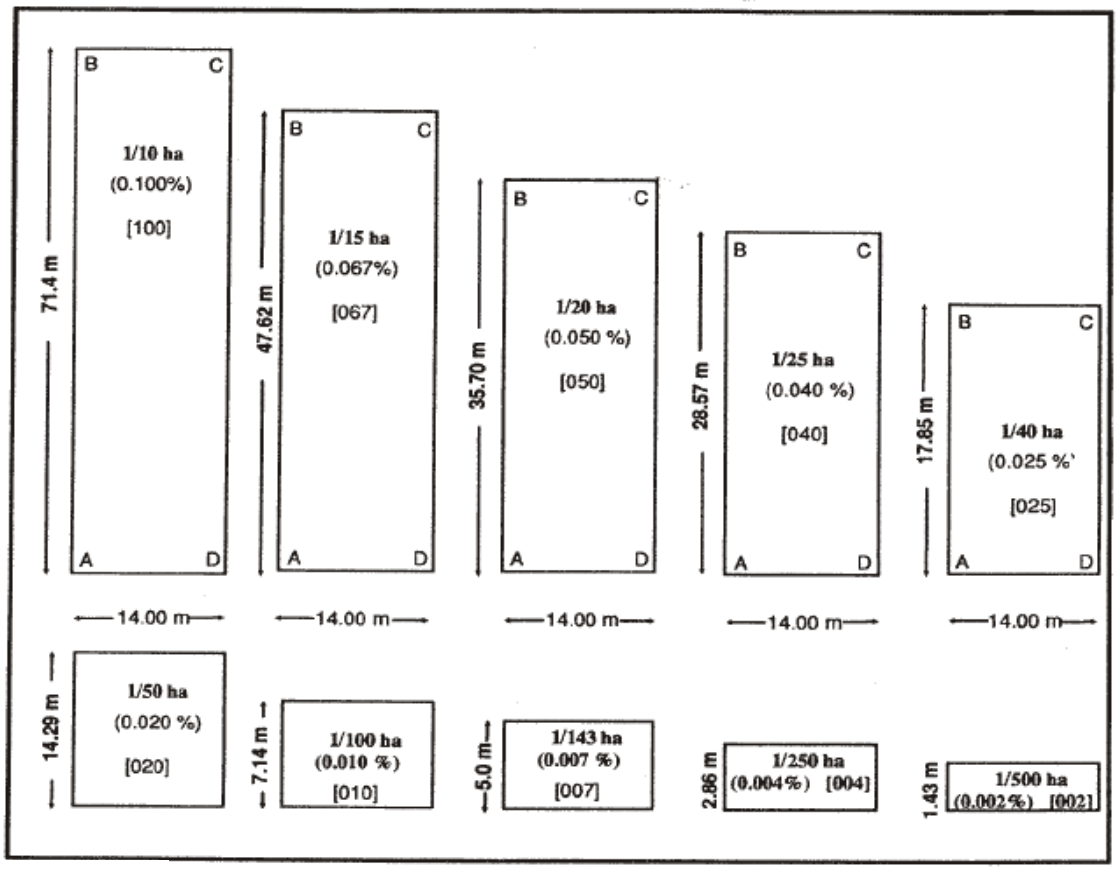


Figure 2. Étendue spatiale de l'acquisition ALS. Système de coordonnées utilisé : UTM 21N.

Le terme placette se définit comme : « Une petite surface de forêt choisie comme étant représentative d'une zone plus étendue »[1]. Un total de 301 placettes dominées par le sapin baumier et l'épinette noire hors de l'étendue d'acquisition des données ALS étaient disponibles pour la validation indépendante du modèle. Les quatre coins et le centre de chaque placette ont été mesurés à l'aide d'un système de positionnement global (GPS) Trimble GeoXH. Toutes les placettes sont de forme rectangulaire, avec une largeur de 14 m. La taille des placettes variait selon la densité de la tige pour les types de peuplements matures et semi-matures et fixés à 0,04 ha pour les types matures et surannés (Figure 3).



Légende

1 / xx ha = Fraction de l'aire en

(0.xxx) = Aire en hectares

[xxx] = Code représentant la taille de chacune des

Figure 3. Taille des échantillons permanents du ministère des Ressources naturelles de Terre-Neuve (van Lier et Luther, Data dictionary, 2017).

[1] Définition tirée du grand dictionnaire terminologique de l'Office québécois de la langue française (http://gdt.oqlf.gouv.qc.ca/)

De nombreux attributs de structure forestière ont été déterminés pour chaque PSP, mais seul le volume total (Tvol) a été utilisé pour cette étude puisqu'il s'agit d'un attribut important du point de vue de la gestion forestière. Le Tvol est une mesure des ressources en bois et un attribut clé de l'inventaire forestier national du Canada (Gillis et al., 2005). Le Tvol a été estimé pour chaque arbre en utilisant les équations spécifiques à l'espèce de Warren et Meades (1986) et de Ker (1974), puis additionné pour chaque placette.

Étant donné que la qualité des données joue un rôle déterminant dans la modélisation, différentes vérifications ont été réalisées sur les données. D'abord, les perturbations de la forêt, et leur niveau de sévérité, survenus entre le jour de la mesure terrain et le jour d'acquisition du satellite Landsat ont été vérifiés, tant au niveau des placettes d'entrainement que de validation. Puisque le nombre de placettes d'entrainement disponibles était peu élevé, seulement celles ayant subi une perturbation d'une sévérité supérieure à 30 % ont été retirées. L'erreur induite dans la construction des modèles était ainsi minimisée tout en permettant de conserver un nombre de placettes acceptable. Ainsi, 61 placettes d'entrainement ont été disponibles pour le développement des modèles de prédiction du Tvol de la stratégie directe et de la phase 1 de la stratégie indirecte (Tableau 1). De ces 61 placettes d'entrainement, 22 ont été déterminées par pixels proxy (donc à partir d'une autre image selon la procédure du BAP). Malgré qu'elles ne provenaient pas de l'image originale, ces 22 pixels ont été conservés étant donné le faible nombre de placettes disponibles dans le but d'assurer une meilleure représentativité. Un total de 177 placettes indépendantes a été disponible pour la validation des modèles. Puisque leur nombre était élevé, de loin supérieur au nombre de placettes d'entrainement, toutes les placettes de validation ayant subi une perturbation d'une sévérité supérieure à 10 % (6 placettes) ont été retirées ainsi que tous les pixels proxy.

Tableau 1. Statistiques descriptives du Tvol des placettes échantillonnées utilisées pour la prédiction.

	Volume total (m^3ha^{-1})			
	Min	Max	Moyenne	Écart-type
Placettes d'entrainement Épinette noire et sapin baumier (n =61)	50,78	320,69	192,99	75,69
Placettes de validation Épinette noire et sapin baumier (n = 177)	1,86	350,55	146,93	69,7

3.2.2 Imagerie satellitaire et données auxiliaires

Une série d'indices ont été dérivés au niveau de la placette et sur une grille à l'échelle de la province pour permettre la prédiction et la cartographie du Tvol (variable réponse) en fonction des facteurs affectant la croissance des arbres. Les bandes spectrales et les indices du produit « Best Available Pixel » (BAP) de Landsat 2011 (White et al., 2014), tels que Vert, NIR, Indice de « Normalized Difference Végetation Index » (NDVI) et l'indice de végétation « Soil-Adjusted Vegetation Index » (SAVI), ont été générés pour représenter la composition et la structure de la végétation. Puisque les bandes d'images satellites sont fortement corrélées, les transformations contribuent généralement à produire des analyses plus efficaces. Une description du produit BAP et de ces indices spectraux dérivés peut être trouvée dans White et al. (2014) et Landsat Surface Reflectance-Derived Spectral Indices product Guide of USGS (2017) (Masek et al., 2006).

Les variables PALSAR 2010, rééchantillonnées au 20 m, incluant la polarisation radar HH et HV, ont également été utilisées comme variables prédictives pour représenter la structure de la forêt (Données téléchargées à ce lien : http://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/data/index.htm). La mosaïque PALSAR, de résolutions globales de 25 m, est une image SAR globalement homogène créée en mosaïquant les images SAR dans des coefficients de rétrodiffusion mesurés par PALSAR, où toutes les bandes de données ayant une latitude et une longitude de 10 x 10 degrés sont traitées par bande et mosaïquées pour des raisons d'efficacité de traitement. Une correction de la distorsion géométrique spécifique au SAR (orthorectification) et les effets topographiques sur l'intensité de l'image (correction de la pente) sont appliqués pour faciliter la classification des forêts. L'intervalle temporel de la mosaïque est généralement d'un an.

Les variables géographiques incluent la latitude, la longitude, tandis que la pente, l'élévation, la SSINA, l'indice de charge thermique 3 (HLI3) (McCune and Keon, 2002) font partie de quelques indices topographiques et de rayonnement solaire calculés. Les variables retenues sont celles largement utilisées dans les recherches sur la modélisation forestière et le suivi de la végétation (Evans et al., 2009 ; Murphy et al., 2010 ; Dech et al., 2014). La série de variables climatiques a été estimée à l'aide de la méthode d'interpolation ANUSPLIN et des mesures météorologiques prises dans les stations météorologiques d'Environnement Canada distribuées sur l'ensemble de la province de Terre-Neuve (McKenney et al., 2007). Les précipitations annuelles (Annual Precipitations: AP), les précipitations totales de la saison de croissance (Total Precipitations during Growing Season:

TPGS), le nombre de jours de croissance (Number of Days during Growing Season: NDGS) et la température moyenne de la saison de croissance (Mean Temperature during Growing Season: MTGS) sont les variables considérées pertinentes pour la productivité forestière. Les variables climatiques énumérées ci-dessus ont été rééchantillonnées en tant que couches matricielles avec une résolution de grille de 20 m × 20 m à partir d'une résolution originale de 150 secondes d'arc (McKenney et al., 2007). Tous les jeux de données ont été rééchantillonnés à une résolution de 20 m selon la méthode bilinéaire de l'outil de rééchantillonnage d'ArcMap 10.5, afin de correspondre aux tailles des plus grandes placettes terrain, et ont été reprojetés en UTM 21N.

Toutes les 28 variables explicatives candidates sont représentées comme une couche raster pour l'ensemble de l'île. Chacune de ces couches a fait l'objet d'une analyse afin de déceler les valeurs aberrantes et les anomalies. Les statistiques descriptives et les histogrammes des données ont été analysés. Ensuite, les données ont été filtrées selon un seuil correspondant à 30 % des valeurs minimales et maximales des placettes indépendantes. Parmi les différents seuils testés, ce dernier était celui optimisant le mieux les résultats. Ensuite, la fonction de R « multi-collinear » a été employée afin d'éliminer les variables présentant un seuil de multicolinéarité supérieur à 0,01. Ce seuil a été utilisé puisque le nombre de variables prédictives était supérieur à 20 et dans le but d'éliminer dès le départ un maximum de variables trop corrélées. Les variables candidates SAVI, MSAVI, NIR TCGRE et TCWET ont été retirées à la suite de cette opération puisqu'elles ne respectaient pas le seuil de multicolinéarité maximum avec les autres variables. Finalement, 23 variables ont été retenues pour le développement des modèles utilisant les données à couverture complète (Tableau 2).

Tableau 2. Variables explicatives candidates à couverture complète conservées pour la prédiction du Tvol.

Type	Nom du prédicteur	Signification
Bandes spectrales Landsat BAP et indices	Blue	
	Green	
	Red	
	SWIR	Shortwave Infrared 1
	SWIR2	Shortwave Infrared 2
	EVI	Enhanced Vegetation Index
	NDMI	Normalized Difference Moisture Index
	NDVI	Normalized Difference Vegetation Index
	TCBRI	Tassel cap brightness
Radar PALSAR	HH	Normalized radar gamma-naught in HH
	HV	Normalized radar gamma-naught in HV
Topographique et indices de radiation solaire	Elevation	
	Slope	Pente
	SCOSA	SCOSA= tan(Slope) ×cos (Aspect)
	SSINA	SSINA= tan(Slope) ×sin (Aspect)
	HLI3	Heat load index third equation
	TWI	Topographic wetness index
Géographique	Longitude	Longitude en degré
	Latitude	Latitude en degré
Climatique	AP	Précipitation annuelle
	NDGS	Nombre de jours de la saison de croissance
	TPGS	Précipitation totale de la saison de croissance
	MTGS	Température moyenne de la saison de croissance

3.2.3 Données ALS

Les données ALS ont été acquises en utilisant le scanneur Optech ALTM 3100C monté sur un avion Piper Navajo, avec une altitude allant de 600 à 1 200 m, durant la saison de croissance en 2010 et 2011. La densité moyenne des impulsions était de 1 à 4 points/m^2, avec jusqu'à quatre retours étant enregistrés par impulsion. L'acquisition ALS visait à couvrir les placettes terrain (PSP) disponibles afin de soutenir le développement du modèle (~1 840 km^2) et d'obtenir un échantillonnage systématique de 12 transects orientés nord-sud, d'environ 500 m de large, espacés de 8 km (~1 060 km^2, soit 1,2 % la zone forestière commerciale) (voir Hopkinson et al., 2013 et la Figure 2 pour plus de détails). Le prétraitement a été effectué conformément aux procédures de calcul de la position du point laser (Milne et al., 2012). Les données ALS calibrées ont ensuite été tuilées, alignées, nettoyées et classées en tant que retour laser terrestre ou non, puis par retour unique, premier, deuxième, troisième et dernier retour et enfin, ajuster au géoïde par le fournisseur de données : Applied Geomatics Research Group de la Nouvelle-Écosse (AGRG).

4. Méthode

Trois grandes étapes ont été nécessaires afin de prédire le volume total de forêt sur l'île de Terre-Neuve (Figure 4). La première étape du projet consistait à produire et sélectionner une série de métriques ALS décrivant le nombre de retours, la hauteur du couvert et sa densité. Les métriques retenues allaient être utilisées comme variables explicatives (indépendantes) dans la phase 1 de la stratégie indirecte pour la prédiction du Tvol. La deuxième étape a été de développer les modèles prédictifs pour chacune des stratégies, soit directe et indirecte, selon les approches paramétriques (OLS) et non-paramétriques (RF) pour chacune des stratégies. Les procédures de sélection des modèles étaient différentes selon la méthode utilisée (OLS ou RF). Ces procédures sont décrites à la section 4.2 ci-dessous. Cependant, le même ensemble de variables prédictives initiales a été utilisé pour les deux méthodes afin de pouvoir comparer les résultats de modélisation. La troisième étape visait l'évaluation des stratégies de modélisation (directe versus indirecte) ainsi que la comparaison des deux méthodes de modélisation utilisées (OLS et RF) pour prédire le Tvol. L'évaluation de la performance des modèles a été évaluée par des statistiques calculées au niveau du développement du modèle et de la prédiction du Tvol sur les placettes de validation indépendantes.

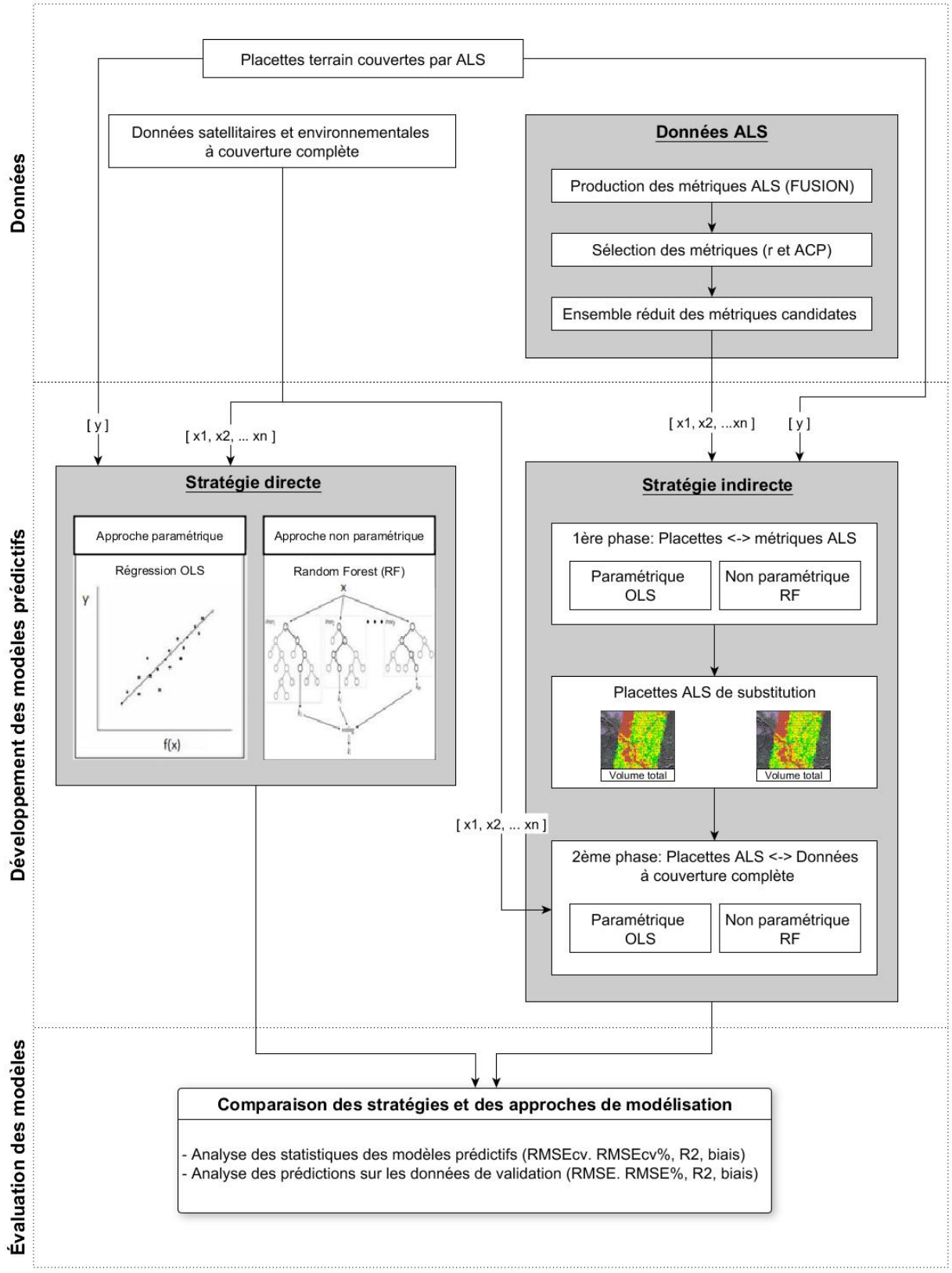


Figure 4. Organigramme des étapes méthodologiques réalisées pour modéliser le Tvol avec la méthode de régression OLS et Random Forest pour les stratégies directes et indirectes.

4.1 Production et sélection des métriques ALS

En règle générale, la plupart des applications forestières nécessiteront des informations sur la hauteur, sur la variabilité de la hauteur et sur la quantité de couvert végétaux présents (Lefsky et al., 2005). Le nuage de points ALS contient des mesures dans un espace tridimensionnel (x, y, z), et des statistiques descriptives peuvent être générées à partir de ces mesures pour résumer le nuage de points de manière statistiquement et spatialement significative (McGaughey, 2014). Ainsi, la première étape du projet consistait à calculer une série de métriques ALS avec le logiciel FUSION, incluant celles reliées au nombre de retours (par ex., retours totaux, par écho, etc.), celles décrivant la hauteur de la canopée (par ex., max, min, 99e percentile, etc.) et celles décrivant la densité de points (par ex., « fractional cover »). Certaines études ont calculé les métriques séparément pour les premiers et derniers retours (par exemple, Næsset 2002) et d'autres ont utilisé tous les retours pour calculer des mesures (par exemple, Woods et al., 2011). Les deux méthodes de calcul des métriques ont produit des modèles prédictifs robustes et, à ce jour, aucune étude dans les revues à comité de lecture n'a examiné de manière rigoureuse si une méthode convenait mieux que l'autre dans les divers environnements forestiers. Les métriques de Terre-Neuve ont été calculées en utilisant tous les retours au niveau de la placette et sur une grille de 20 m × 20 m. Plusieurs études utilisant l'approche par aire définie (area-based approach: ABA) ont appliqué un seuil hauteur minimale de 2 m pour le calcul des métriques (par ex., Næsset, 2002 ; Andersen et al., 2005 ; Frazer et al., 2011 ; Hyyppä et al., 2012 ; Wulder et al., 2012). Les recherches ont démontré que ce seuil est approprié dans des conditions de forêt boréale possédant des arbres matures. Par conséquent, ce seuil de hauteur minimale de 2 m a été appliqué pour le calcul des métriques dans la présente étude sur l'île de Terre-Neuve. Une hauteur de rupture de 3 m a été utilisée pour calculer les métriques d'estimation de la couverture forestière (ratio des retours au-dessus d'une hauteur déterminée).

Avant de procéder à la sélection des métriques candidates pour la prédiction du Tvol, un prétraitement a été réalisé pour s'assurer de la qualité des données. D'abord, seuls les pixels ayant un nombre de premiers retours supérieur à 300 et une élévation supérieure à 2 m ont été retenus. Ensuite, les statistiques descriptives des métriques produites ont été analysées pour rechercher les valeurs anormales et les valeurs aberrantes. Dans l'optique de prévoir l'échantillonnage avant la phase 2 de la stratégie indirecte, les pixels n'étant pas considérés comme de la forêt ont été enlevés. Les pixels non-forestiers ont été exclus en fonction d'un masque créé à partir des polygones forestiers ayant été délimités par photo-interprétation de l'imagerie aérienne.

Les premières métriques à avoir été retirées d'emblée des variables candidates pour la modélisation sont celles du compte de retour total, car la densité des points était très variable sur l'ensemble du territoire (35 à 1480). Les métriques d'intensité ont aussi été retirées. Toutefois, les métriques de ratio ont été conservées. Il n'en demeurait pas moins que plus d'une cinquantaine de métriques étaient toujours possibles et plusieurs étaient hautement corrélées. Conséquemment, deux approches ont été utilisées afin de réduire l'ensemble des métriques candidates pour la prédiction du Tvol. D'abord, la corrélation de Pearson (r) a été utilisée pour identifier les variables explicatives hautement corrélées ($r > 0,9$), comme présentée dans Hudak et al. (2012) et Silva et al. (2014), afin de réduire l'ensemble des variables explicatives candidates pour la prédiction du Tvol. Si un groupe donné (deux ou plusieurs) de métriques ALS était fortement corrélé, une seule mesure a été retenue, en excluant celles qui étaient le plus fortement corrélées avec les métriques restantes. Ensuite, une analyse des composantes principales (ACP) a été effectuée avec les métriques retenues. Généralement, lorsque l'ACP est réalisée avec des variables normalisées, les valeurs propres des composantes principales (CPs) sont utilisées pour déterminer le nombre d'axes principaux à conserver. Une valeur propre > 1 indique que la composante principale concernée représente plus de variances par rapport à une seule variable d'origine. Ceci est généralement utilisé comme seuil à partir duquel les composantes principales sont conservées (Kaiser, 1961). Puisque le nombre de composantes principales à conserver peut dépendre du domaine d'application et du jeu de données spécifiques, la décision a aussi été basée sur le graphique des valeurs propres (appelé scree plot). Le nombre d'axes à conserver est déterminé par le point au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables (Jollife, 2002 ; Peres-Neto et al., 2005). Ensuite, l'observation des vecteurs propres (eigenvectors) de chacune des métriques des composantes principales conservées a permis de révéler les métriques qui étaient le plus susceptibles de contribuer au développement des modèles. Les métriques ayant la plus haute contribution (loading eigenvectors) dans chacune des composantes principales retenues ont été conservées comme variables prédictives pour le développement des modèles de la phase 1 de la stratégie indirecte permettant de prédire le Tvol.

4.2 Développement des modèles prédictifs

Les performances des méthodes paramétriques (OLS) et non-paramétriques (RF) ont été analysées et comparées pour les stratégies directes et la stratégie indirecte. Cette section décrit les principes propres à chacune des stratégies de modélisation et la manière dont les approches paramétriques (OLS) et non-paramétriques (RF) ont été implémentées au sein de chacune des stratégies de modélisation. Chaque méthode de modélisation (OLS et RF) est implémentée selon les mêmes variables sélectionnées et les mêmes calculs statistiques ont été utilisés pour l'évaluation de la performance des modèles et la précision de la prédiction du Tvol pour chacune des stratégies.

4.2.1 Stratégie directe

Pour la stratégie directe, des relations ont été développées entre les mesures disponibles des placettes de terrain et les variables disponibles en couverture complète pour l'ensemble de la zone d'intérêt, c'est-à-dire l'île de Terre-Neuve. Ces variables étaient des données environnementales ou climatiques, de l'information géographique (par ex., topographique) ou des indices spectraux (Tableau 2). Les relations établies ont ensuite été utilisées afin de prédire le Tvol des placettes de validation dans l'optique d'évaluer la précision de la stratégie directe.

4.2.2 Stratégie indirecte

La stratégie indirecte comprenait deux étapes et combine des placettes forestières, des transects ALS et des couches de données spatialement complètes pour toute l'île de Terre-Neuve afin de prédire le Tvol. Dans la première étape, les relations ont été modélisées entre les données ALS et les mesures des placettes terrain. Dans cette étape, la prédiction de l'attribut forestier Tvol était spatialement limitée à la zone où les données ALS étaient collectées. La méthode d'échantillonnage « Local Pivotal Spatially Balanced » pour de grands échantillons (lpm_kdtree) (Grafström et Tillé, 2013), de la librairie « SamplingBigData » du logiciel R (R Development Core Team; Lisic et Grafström, 2018), a été utilisée pour échantillonner les placettes de substitution nécessaires à la deuxième étape de modélisation. La structure de données k-d tree utilisée pour implémenter la méthode « Local Pivotal » permet de réduire le temps d'exécution de l'échantillonnage.

La définition des probabilités d'inclusion a été prescrite en fonction de la population finie à l'intérieur des 10 strates appliquées sur les valeurs de Tvol (Figure 5). Le paramètre d'algorithme de recherche du plus proche voisin a été défini par « kdtree ». Cette méthode a l'avantage de sélectionner des

échantillons avec un haut degré d'équilibre spatial bien répartis dans l'espace auxiliaire. En d'autres mots, cela permet de sélectionner des échantillons distribués sur l'étendue des valeurs de la population d'intérêt ainsi que représentant la distribution des valeurs des variables explicatives. Cinq variables auxiliaires ayant la plus forte corrélation avec la variable d'intérêt (Tvol) ont été sélectionnées comme variables d'équilibrage pour l'échantillonnage dans les données ALS.

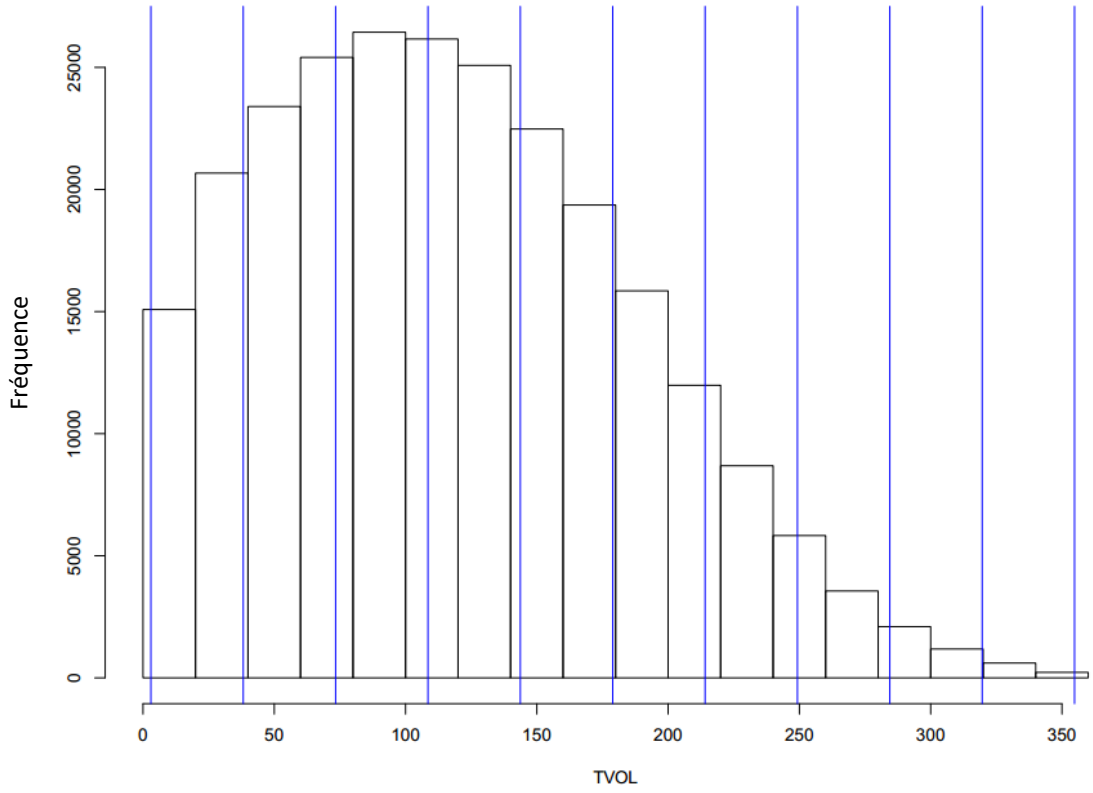


Figure 5. Représentation de la distribution du Tvol des placettes de substitution potentielles en dix strates de même étendue, à partir des données ALS pour la phase 2 de la stratégie indirecte de prédiction du Tvol.

Les placettes ALS de substitution potentielles (pixels couverts par les données ALS avec les valeurs de Tvol prédites à la première étape) ont été limitées aux pixels correspondants à de la forêt et à la gamme des valeurs de Tvol mesurées dans les placettes forestières de Terre-Neuve afin d'obtenir des données d'entrainement représentatives des conditions forestières. Les pixels qui ne provenaient pas du produit Landsat BAP 2011 original, c'est-à-dire qu'ils provenaient d'un pixel proxy, n'ont pas non plus été retenus en tant que placettes ALS de substitution potentielles. Dans la deuxième étape, les relations ont été modélisées entre les placettes de substitution échantillonnées à l'intérieur des transects ALS (Tvol prédit lors de la phase 1) et les variables prédictives à couverture complète, telles

que les bandes spectrales de l'imagerie optique, les données topographiques et climatiques. La prédiction de Tvol a ensuite été étendue de la zone couverte par la donnée ALS à l'ensemble du territoire. À ce stade, les statistiques d'erreurs des modèles ont été produites. De plus, la performance du modèle a été évaluée avec les placettes de validation comme pour la stratégie directe.

4.2.3 Prédiction du volume total avec la méthode OLS

La méthode OLS a été choisie comme approche de modélisation paramétrique à mettre en œuvre dans les stratégies de modélisation directe et indirecte de la prédiction de l'attribut forestier Tvol. La sélection des modèles candidats a été effectuée à l'aide de la technique « best subset regression » disponible dans la librairie « Leaps » du logiciel R (R Development Core Team, Lumley, 2017). Leaps utilise un algorithme « branch-and-bound » efficace pour effectuer une recherche exhaustive des meilleurs sous-ensembles de variables explicatives de la variable dépendante dans la régression linéaire (Hudak et al., 2006). La méthode statistique du Cp de Mallows (Mallows, 1973) (Équation 1) a été choisie pour identifier les 10 meilleurs modèles de chaque combinaison de prédicteurs.

$$Cp = \frac{SS_{res}}{MS_{res}} - N + 2p \qquad (1)$$

Dans l'équation du Cp de Mallows, le SSres correspond à la somme des carrés résiduels pour le modèle avec les variables p-1, MSres est le carré moyen résiduel lors de l'utilisation de toutes les variables disponibles, N est le nombre d'observations et p est le nombre de variables utilisées pour le modèle plus un. La statistique du Cp de Mallows est un critère servant à évaluer l'ajustement du modèle quand des modèles avec un nombre différent de paramètres sont à comparer. Il compare la précision et le biais du modèle complet à ceux des modèles contenant un sous-ensemble de prédicteurs. Il s'agit donc d'un indicateur de l'équilibre entre les modèles qui sont trop simples, qui peuvent souffrir de coefficients biaisés et de prévisions biaisées, ou trop compliquées, ce qui entraine une variance importante dans les coefficients et la prédiction (Myers, 1990). Le modèle ayant le Cp de Mallows le plus petit et le plus proche du nombre de variables explicatives, additionné à la constante (p), dans le modèle est à privilégier. Une petite valeur de Cp de Mallows indique que le modèle estime avec une précision relative (petite variance) les véritables coefficients de régression et les prédictions. Lorsque la valeur de Cp de Mallows est proche du nombre de prédicteurs et de la constante (p), ceci indique que le modèle est relativement sans biais. À l'inverse, les modèles

ceprésentant un manque d'ajustement et des biais ont des valeurs de Cp de Mallows plus grands que la constante (p).

Un certain nombre de tests statistiques ont été effectués pour s'assurer que le modèle paramétrique pouvait fonctionner dans les limites des hypothèses de base de la régression linéaire. La normalité des résidus a été évaluée par les tests de Shapiro-Wilk (Royston, 1982) et l'homogénéité de la variance par le test de Breusch-Pagan (Breusch et Pagan, 1979). L'indice de « Variance Inflation Factor » (VIF) a été calculé afin d'identifier les problèmes de multicolinéarité (Myers, 1990). Chaque modèle présentant une hétéroscédasticité, une non-normalité des résidus ($p < 0,05$) ou une colinéarité ($VIF > 10$) a été automatiquement retiré de l'ensemble des modèles candidats.

Une fois les modèles candidats mis en place, les critères d'information d'Akaike (Akaike Information Criteria : AIC) (Akaike, 1973) et ses mesures associées, soit le Delta d'Akaike (Δ_i) et le poids d'Akaike (ω_i), ainsi que le critère d'information bayésien (Baysian Information Criteria : BIC) de Schwarz (1978) ont été utilisés pour sélectionner le meilleur modèle pour la prédiction du Tvol. L'AIC permet d'identifier les modèles les plus parcimonieux, soit les modèles avec un minimum de biais (qui diminue avec le nombre de paramètres) et un maximum de précision (qui augmente en décrivant les données avec le plus petit nombre de paramètres) (Mazerolle, 2006 ; Symonds and Moussalli, 2011) (Figure 6). Le critère du BIC (Schwarz, 1978) pénalise davantage le nombre de variables présentes dans le modèle que le critère d'AIC. Il vise la sélection de variables statistiquement significative alors que l'AIC tente de retenir des variables pertinentes lors de prévisions.

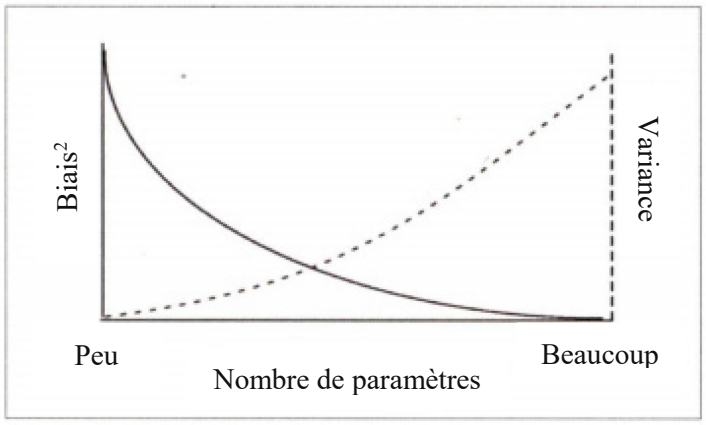


Figure 6. Principe de parcimonie pour la mise en place de modèles (Posada, 2004).

Le calcul de l'AIC (Équation 2) pour chaque modèle candidat a été effectué à l'aide de leurs vraisemblances et du nombre de paramètres (K) qui a été nécessaire d'estimer pour les développer. Plus particulièrement, le paramètre K correspond au nombre de variables incluses dans le modèle ainsi que l'estimation de l'ordonnée à l'origine et du paramètre associé à l'estimation de l'erreur. L'intégration du paramètre K dans l'équation des AIC permet de pénaliser les modèles qui ont trop de variables, soit les modèles moins parcimonieux. Les modèles candidats de la stratégie directe et de la première phase de la stratégie indirecte ont été classés en fonction de l'AIC corrigée pour la petite taille de l'échantillon (AICc). L'AICc (Équation 3) intègre aussi le nombre d'échantillons (n) ce qui a pour effet de pénaliser davantage les modèles nécessitant l'estimation d'un trop grand nombre de paramètres :

$$AIC_i = -2 * (\log(vraisemblance_i)) + 2K_i \quad (2)$$

$$AICc_i = AIC + \frac{2K(K+1)}{n-K-1} \quad (3)$$

Le ω_i a ensuite été calculé afin de comparer entre eux les modèles (Équation 4). Le ω_i est une normalisation sur une échelle de 1 du Δ_i (Équation 5) soit la différence entre le AICc de chaque modèle et celui considéré comme le meilleur (minAICc). Le ω_i obtenu peut être interprété comme le pourcentage de chance que le modèle associé soit le plus parcimonieux parmi l'ensemble des modèles candidats (Mazerolle, 2006). Dans tous les cas, le modèle associé au plus grand ω_i est celui avec l'AICc le plus faible :

$$\omega_i = \frac{\text{Exp}\left(-\frac{\Delta_i}{2}\right)}{\sum_{k=1}^{R} \text{Exp}\left(-\frac{\Delta_i}{2}\right)} \quad (4)$$

$$\Delta_i = AICc_i - \min(AICc) \quad (5)$$

De plus, le coefficient de détermination ajusté pour le nombre de prédicteurs (R^2_{adj}) et l'erreur quadratique moyenne validée de manière croisée (RMSEcv) (Myers, 1990 ; Ohmann et Gregory,

2002) ont été calculés à l'aide du logiciel R dans le but d'évaluer la variance expliquée par le modèle et la précision de la prédiction. En ce qui concerne le R^2 ajusté, les modèles candidats sont évalués en fonction de la variance expliquée par les coefficients de détermination ajustés pour le nombre de prédicteurs. C'est une mesure qui permet, comme le R^2, d'évaluer le degré d'adéquation du modèle. Les valeurs possibles sont comprises entre 0 et 1. Le pouvoir explicatif du modèle est fort lorsque R^2 s'approche de la valeur 1. L'avantage majeur du R^2 ajusté est qu'il n'augmente pas avec l'introduction de nouvelles variables peu corrélées avec la variable dépendante (Y) dans le modèle (Dodge, 1999).

4.2.4 Prédiction du volume total avec la méthode Random Forest

RF a été choisi comme approche de modélisation non-paramétrique à mettre en œuvre dans l'approche directe et indirecte de la cartographie de l'attribut forestier Tvol. L'implémentation de RF a été réalisée avec la librairie « ModelMap » disponible sur le logiciel R (R Development Core Team, Freeman et al., 2009). Deux paramètres principaux doivent être définis par l'utilisateur pour la modélisation avec RF ; le nombre de variables candidates sélectionnées sur chaque groupe de nœuds (mtry) et le nombre total d'arbres construits dans chaque forêt (ntree). La valeur de mtry à utiliser a été déterminée en testant le mtry par défaut, puis avec la moitié de cette valeur et le double. La valeur de mtry par défaut est calculée en divisant par trois le nombre total de variables de prédiction (Liaw and Wiener, 2013). Dans ce projet, la valeur de mtry par défaut correspondait à 8.

La stabilisation des erreurs « out-of-bag » (l'erreur calculée [Mean Square Error : MSE] à l'aide de l'échantillon retenu de la construction de chaque arbre de régression) s'est produite à environ 500 arbres dans le cas de la stratégie directe et de la phase 1 de la stratégie indirecte et a donc été utilisée comme valeur du paramètre ntree () (Figure 7).

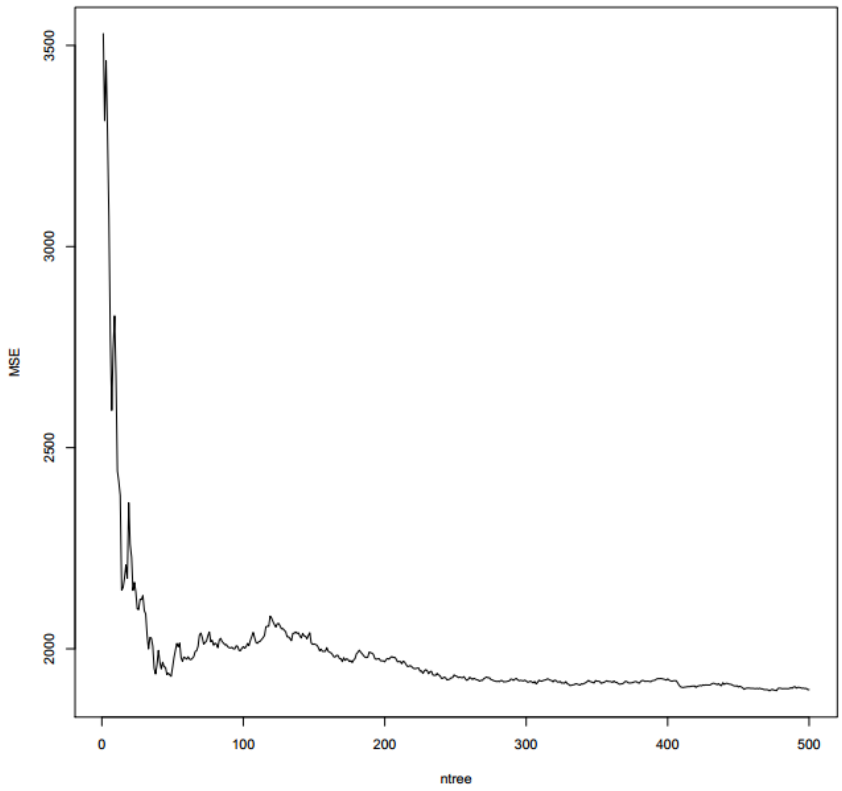


Figure 7. Erreurs de prédiction « out-of-bag » en fonction du nombre d'arbres de décision de RF de la phase 1 de la stratégie indirecte.

4.3 Évaluation des stratégies de modélisation : directe/indirecte et OLS/R F

Plusieurs statistiques ont été calculées au niveau du développement du modèle et des données de validation indépendantes afin de comparer les résultats de chacune des stratégies (directe et indirecte) et de chacune des approches de modélisation (OLF et RF). Le RMSE avec validation croisée de type « *k-fold* » (RMSEcv) (Équation 6) et le RMSEcv (Équation 7) relatif ont été calculés pour évaluer les prédictions réalisées lors de la phase de développement des modèles :

$$RMSEcv = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,-i})^2}{n}} \qquad (6)$$

$$\% \text{ RMSEcv} = \frac{RMSE_{cv}}{\bar{Y}_i} x100 \quad (7)$$

Le RMSE et le RMSE relatif ont été utilisés pour mesurer la précision de la prédiction du Tvol sur le jeu de données de validation indépendant. Le R^2 a également été calculé dans le but d'évaluer la quantité de variances expliquée par le modèle et la précision de la prédiction (Myers, 1990 ; Ohmann et Gregory, 2002). Aussi, le biais des prédictions (Équation 8) a été calculé tant au niveau du développement des modèles que de la validation sur les données indépendantes :

$$\text{Biais} = \frac{\sum \text{Valeur observée} - \text{Valeur prédite}}{\text{Nombre d'observation}} \quad (8)$$

Enfin, des graphiques illustrant les valeurs observées de Tvol versus les valeurs de Tvol prédites modélisées avec les stratégies directes et indirectes ont également été produits. Ces graphiques ont permis une évaluation visuelle de la capacité des modèles à prédire le Tvol.

5. Résultats

5.1 Sélection des métriques ALS

Suite à la production d'une série de métriques ALS à l'aide du logiciel FUSION, il a été nécessaire de réduire le nombre de métriques candidates pour le développement des modèles de la phase 1 de la stratégie indirecte. L'application de deux techniques a permis d'atteindre l'objectif de réduction des variables des métriques ALS, soit la corrélation de Pearson (r) et l'ACP. D'abord, le test de la corrélation de Pearson (r) a démontré que seulement 12 métriques générées par le logiciel FUSION présentaient une corrélation inférieure à 0,9. Le niveau de corrélation entre les différentes variables est présenté à la Figure 8. Le niveau d'intensité de la couleur indique la force de la corrélation. Le chiffre, indiquant la valeur de la corrélation entre les variables, est rouge lorsque la corrélation est négative et bleue lorsqu'elle est positive. La métrique du 99e percentile de hauteur a été conservée malgré sa corrélation supérieure à 0,9 avec les autres, car cette mesure est fréquemment utilisée comme prédicteur potentiel pour la prédiction des attributs forestiers dans d'autres études (Næsset, 2002 ; 2004 ; Garcia et al., 2010 ; Hudak et al., 2012 ; Silva et al., 2014). Ainsi, un total de 13 métriques ALS ont été sélectionnées de l'ensemble original des métriques ALS générées et analysées par ACP (Tableau 3).

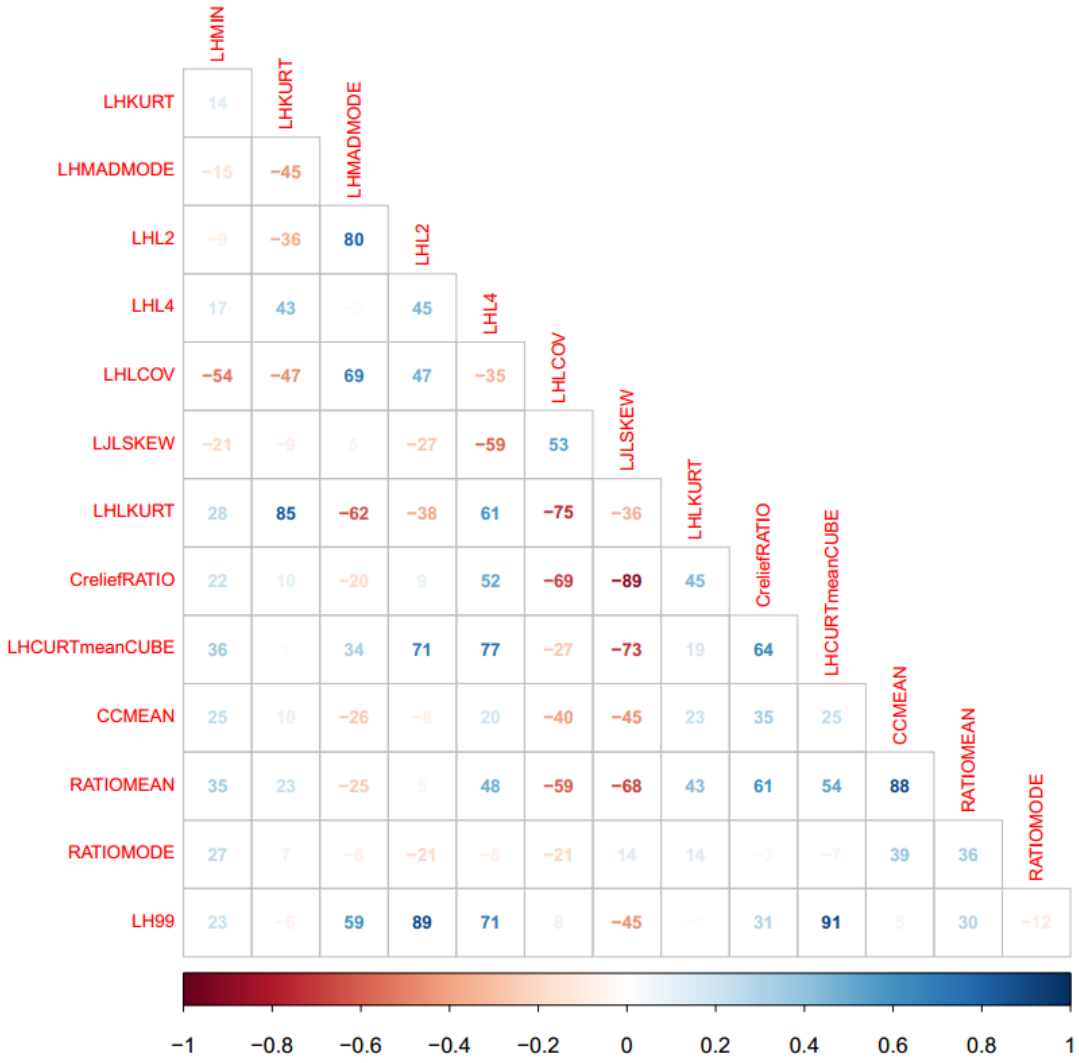


Figure 8. Corrélation des Pearson (r) entre les treize métriques ALS sélectionnées.

Tableau 3. Métriques sélectionnées pour le modèle prédictif de phase 1 de la stratégie indirecte

Nom des métriques	**Définition**
LHMIN	Elevation minimum
LHKURT	Elevation kurtosis
LHMADMODE	Elevation median of the absolute deviations from the overall mode (MAD)
LHL2	Elevation L2-moment
LHL4	Elevation L4-moment
LHLCOV	Elevation L CV
LJLSKEW	Elevation L-moment skewness
LHLKURT	Elevation L-moment kurtosis
CreliefRATIO	Canopy relief ratio
LHCURTmeanCUBE	Elevation CURT mean CUBE
CCMEAN	Percentage of all returns above mean
RATIOMEAN	Number of returns above mean height/Total first returns *100
RATIOMODE	Number of returns above mode height/Total first returns *100
LH99	99 th percentile

La deuxième étape de la sélection des métriques ALS consistait à procéder à une ACP des 13 métriques qui présentaient une corrélation inférieure au seuil déterminé. L'ACP révèle que seules les trois premières composantes principales possèdent des proportions de valeurs propres supérieures à 10 % (valeur propre > 1), et qu'elles expliquent seulement 77,44 % de la variation (Figure 9). La première composante principale explique 38 % de la variance, alors que la deuxième et troisième expliquent 27,6 % et 11,8 % respectivement. Le graphique des valeurs propres révèle qu'il faut considérer un total de sept composantes principales pour atteindre des valeurs relativement de petite taille et comparables. Plus spécifiquement, il faut conserver la CP4, CP5, CP6, CP7 qui expliquent respectivement, 8,5 %, 6,6 %, 3,6 % et enfin 2,1 % de la variance de l'ensemble des métriques ALS.

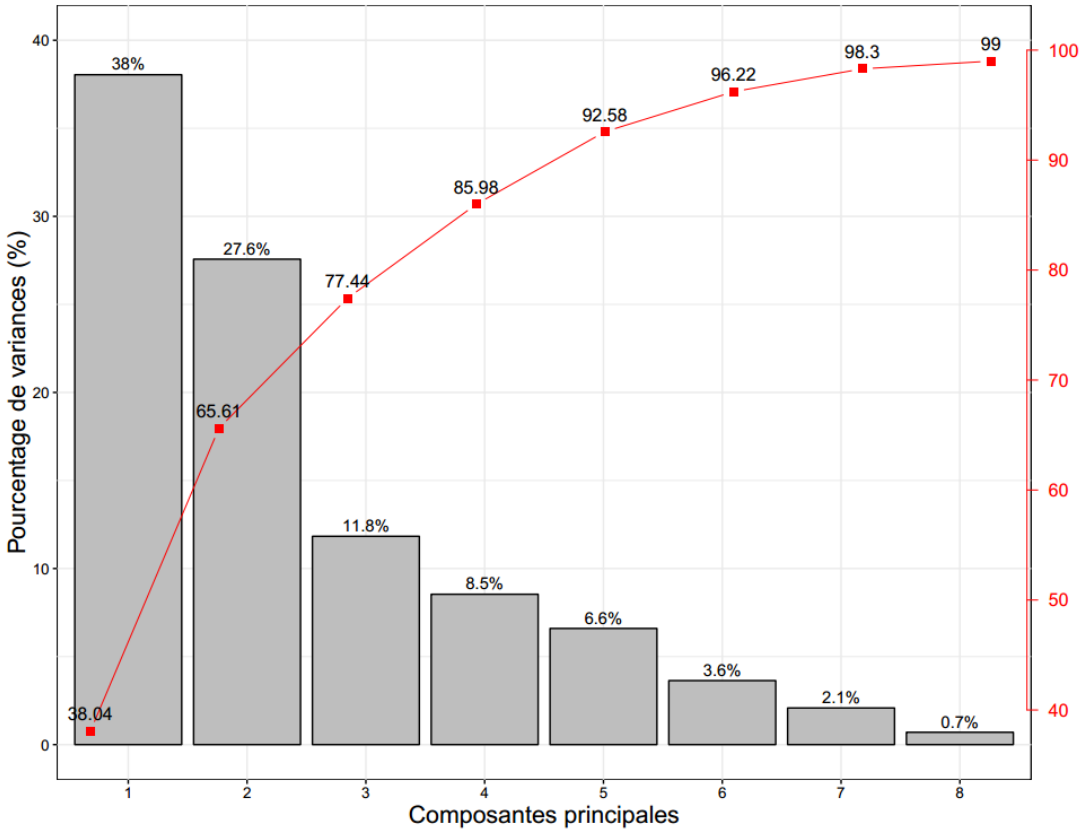


Figure 9. Pourcentage de variance et pourcentage cumulatif de variance de Tvol expliqués par huit composantes principales.

Le graphique de corrélation des variables de l'ACP démontre qu'une majorité des variables sont bien représentées par les CPs 1 et 2 (Dim 1 et Dim 2 de la Figure 10). En effet, 10 des 13 métriques ALS sélectionnées se rapprochent fortement du cercle de corrélation de façon positive ou négative. Les variables les moins corrélées aux deux premières composantes principales sont RATIOMODE, LHMIN et CCMEAN. Selon le calcul de contributions des vecteurs propres (Tableau 4), la métrique RATIOMEAN exprime une contribution positive sur la CP1 alors que LJLSKEW exprime une contribution identique négative (0,36). Cependant, 5 autres métriques possèdent des contributions similaires, soit LHL4, LHLCOV, LHLKURT, CreliefRATIO, LHCURTmeanCUBE. Donc toutes ces variables ont été conservées pour la modélisation du Tvol. Les CP2, CP3 et CP4 démontrent des contributions positives de LHL2, CCMEAN et RATIONMODE respectivement. La métrique LHMIN contribue fortement de manière négative à la CP5. Le RATIOMODE contribue cette fois de manière négative à la CP6 et enfin, la variable LHMADMODE est celle qui présente une contribution

positive à la CP7. Les variables LH99 et LHKURT ont été également conservées comme prédicteurs potentiels puisque leur contribution aux CP2 et CP7 respectivement était importante. La variable LH99 faisait également partie des 3 premières variables d'importance des CP1 et CP2 (Figure 11). En somme, toutes les métriques sélectionnées au départ grâce à la corrélation de Pearson ($r > 0,9$) ont finalement été retenues comme variables prédictives potentielles du Tvol.

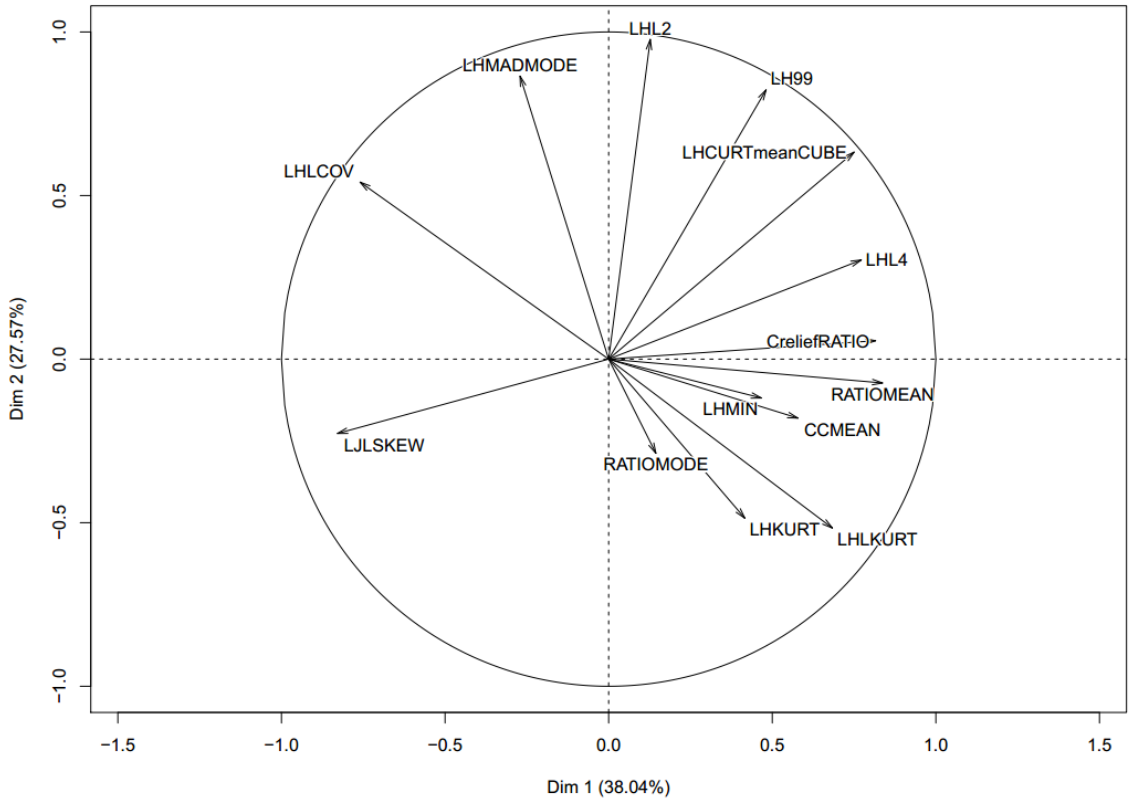


Figure 10. Graphique de corrélation des variables de l'ACP pour les composantes principales CP1 (Dim 1) et CP2 (Dim 2).

Tableau 4. Contribution des vecteurs propres des variables pour les 7 composantes principales conservées.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
LHMIN	0,2	-0,06	0,18	0,35	**-0,73**	0,34	0,09
LHKURT	0,18	-0,25	-0,41	0,36	0,25	0,11	0,53
LHMADMODE	-0,12	0,44	0,09	0,17	0,01	-0,16	**0,6**
LHL2	0,06	**0,5**	-0,02	0,09	0,1	0,02	-0,11
LHL4	0,33	0,15	-0,31	0,18	0,2	0	-0,45
LHLCOV	-0,33	0,28	0,01	0,09	0,3	0,13	0,06
LJLSKEW	**-0,36**	-0,12	-0,02	0,4	0	0,08	-0,29
LHLKURT	0,3	-0,26	-0,34	0,19	0,1	-0,12	0,01
CreliefRATIO	0,35	0,03	0	-0,41	-0,13	-0,41	0,13
LHCURTmeanCUBE	0,33	0,32	-0,02	0,05	-0,13	-0,02	-0,06
CCMEAN	0,25	-0,09	**0,49**	-0,04	0,38	0,43	0,05
RATIOMEAN	**0,36**	-0,04	0,34	-0,01	0,26	0,19	0,03
RATIOMODE	0,06	-0,15	0,47	**0,5**	0,11	**-0,65**	-0,09
LH99	0,21	0,42	-0,08	0,23	-0,04	0,06	-0,1

Note : Consultez le Tableau 3 pour obtenir une description des métriques ALS. Les caractères gras indiquent les métriques avec la plus forte charge sur la composante principale et la plus importante pour une composante principale donnée.

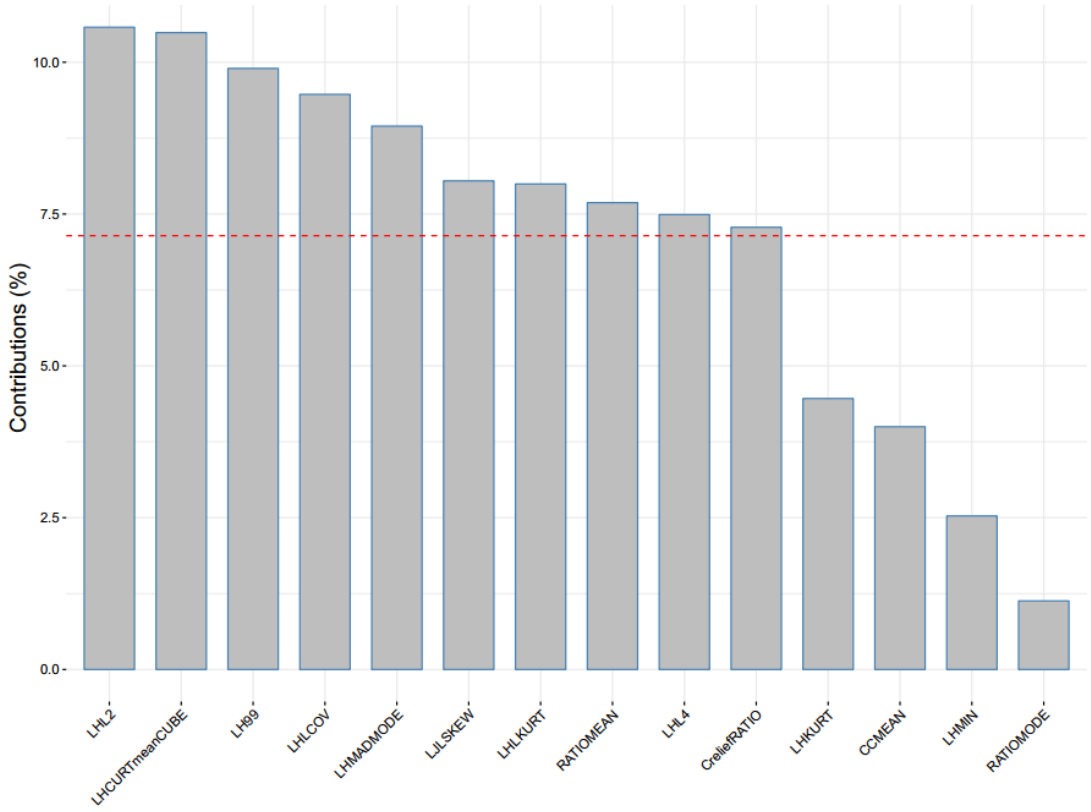


Figure 11. Pourcentage de contribution des variables des composantes principales CP1 et CP2.

5.2 Stratégie directe

À ce stade, la préparation des données a été terminée et la sélection des métriques ALS nécessaires au développement des modèles de la phase 1 de la stratégie indirecte a été réalisée. L'étape suivante consistait donc à développer les différents modèles de la stratégie directe pour prédire le Tvol, soit les modèles OLS et RF.

5.2.1 Régression OLS

La comparaison des 7 modèles utilisant la régression OLS pour la stratégie directe, qui respectait toutes les hypothèses de base des modèles linéaires, à partir des AICs et du poids d'AIC a révélé que le meilleur modèle possédait 4 variables prédictives (Tableau 5). Deux des variables sont des bandes spectrales de Landsat BAP, soit la bande bleue et la bande rouge. Les autres sont le NDVI et la polarisation HH du radar Palsar. Ce modèle explique 44 % de la variance du Tvol et obtient un RMSEcv relatif de 32,07 % lorsque les outils de diagnostic de modèle sont utilisés (Tableau 6). Le graphique des valeurs observées versus les valeurs de Tvol prédites pour le diagnostic du modèle

présente une difficulté de prédiction du Tvol pour l'épinette noire (black spruce : bS) (Figure 12). En effet, les valeurs de Tvol de cette espèce sont éloignées de la ligne du 1:1 dans la majorité des prédictions. Pour le sapin baumier (balsam fir : bF) on observe un écart accentué du Tvol observé versus prédit pour certaines placettes, mais dans une moindre mesure que pour le bS. Lorsque ce modèle est utilisé pour prédire le Tvol sur les 177 placettes de données de validation, la variance expliquée par le modèle chute à 10 % et le RMSE relatif est de 55,59 % (Tableau 7). Le biais devient significatif, avec une valeur négative de -35,11. La droite de régression entre le Tvol prédit et observé a une pente différente de 1, et l'ordonnée à l'origine diffère significativement de 0 pour approcher 60 (Figure 13). Le modèle démontre une tendance à sous-estimer le Tvol pour les deux espèces (bF et bS).

Tableau 5. Sommaire des prédicteurs pour les modèles OLS pour la stratégie directe et indirecte.

Stratégie directe	**Stratégie indirecte**	
Tvol ~ Blue + Red + NDVI + HH	**Phase 1**	Tvol ~ LHKURT + LHLKURT + LHCURTmeanCUBE + CCMEAN
	Phase 2	Tvol ~ Blue + Green + SWIR2 + NDMI + HV + Elevation + Longitude + Latitude

Tableau 6. Sommaire des résultats du diagnostic des modèles pour la stratégie directe et indirecte.

Stratégie directe

		OLS				**RF**		
n	**R^2**	**RMSEcv**	**RMSEcv%**	**Biais**	**R^2**	**RMSEcv**	**RMSEcv%**	**Biais**
61	0,44	61,89	32,07	-5,9 E-13	0,19	67,42	34,93	-0,12

Stratégie indirecte

		OLS				**RF**		
	n	**R^2**	**RMSEcv**	**RMSEcv %**	**Biais**	**R^2**	**RMSEcv**	**RMSEcv %**	**Biais**
Phase 1	61	0,79	38,15	19,77	-2,6 E-13	0,68	43,39	22,48	1,41
Phase 2	5000	0,32	82,80	46,63	-8,8 E-14	0,55	42,56	22,52	1,09

Diagnostics du développement des modèles OLS — Tvol (m^3/ha)

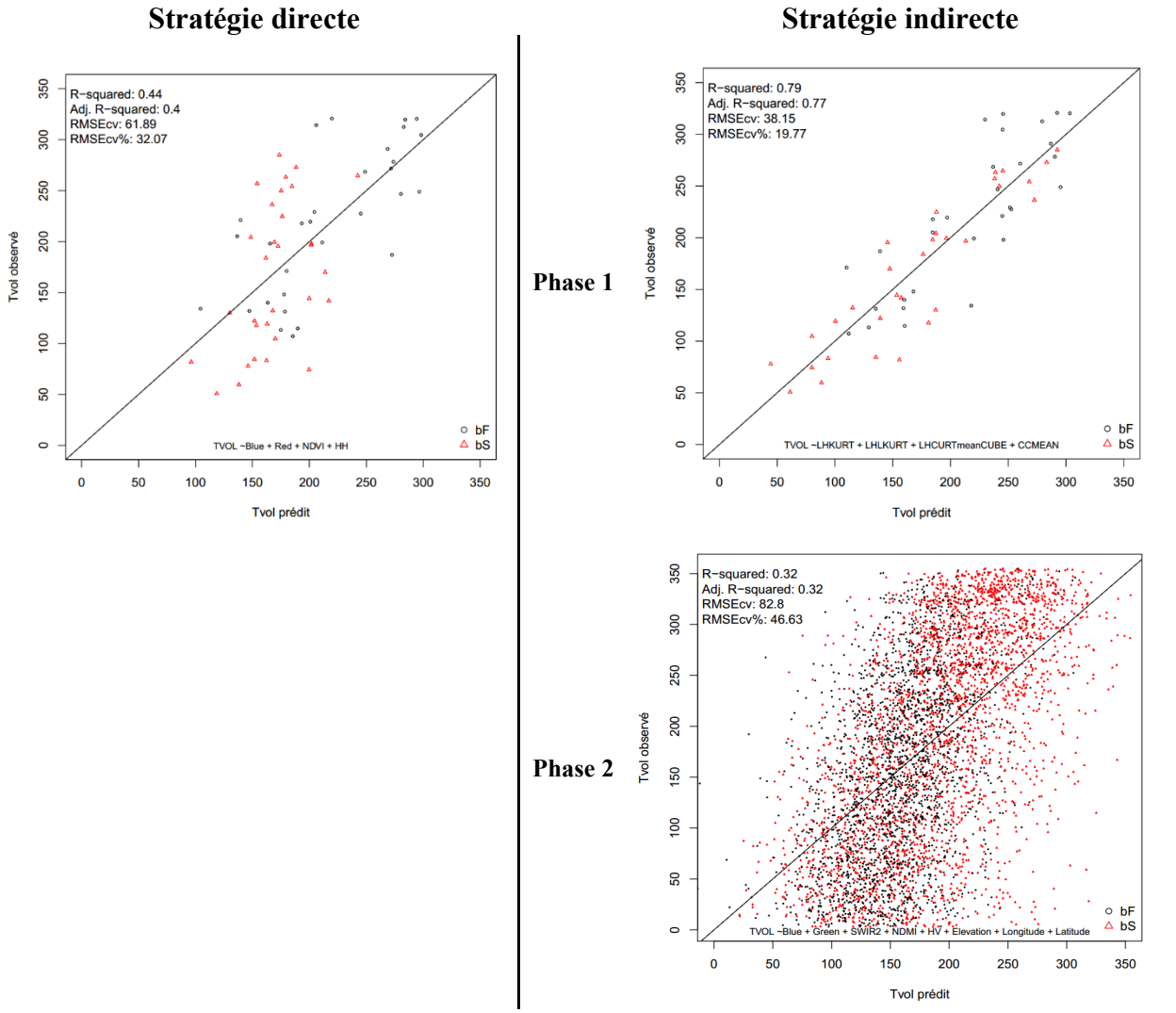


Figure 12. Ligne noire : Valeurs de Tvol observées versus prédites (m^3/ha) modélisées par la régression OLS en utilisant la stratégie directe et indirecte (phase 1 et phase 2), en fonction de la validation croisée « k-fold » pour le diagnostic du développement des modèles Ligne pointillée : Pente = 1 et ordonnée à l'origine = [0:0].

Tableau 7. Sommaire des résultats obtenus à partir des données de validation pour la stratégie directe et indirecte.

Stratégie directe

	OLS				RF			
n	**R^2**	**RMSE**	**RMSE %**	**Biais**	**R^2**	**RMSE**	**RMSE %**	**Biais**
177	0,1	81,58	55,59	-35,11	0,14	72,63	49,48	-32,69

Stratégie indirecte

	OLS				RF			
n	**R^2**	**RMSE**	**RMSE %**	**Biais**	**R^2**	**RMSE**	**RMSE %**	**Biais**
177	0,16	68,94	46,97	-18,01	0,11	72,71	49,54	-30,72

Validation des modèles OLS (177 placettes indépendantes) — Tvol (m^3/ha)

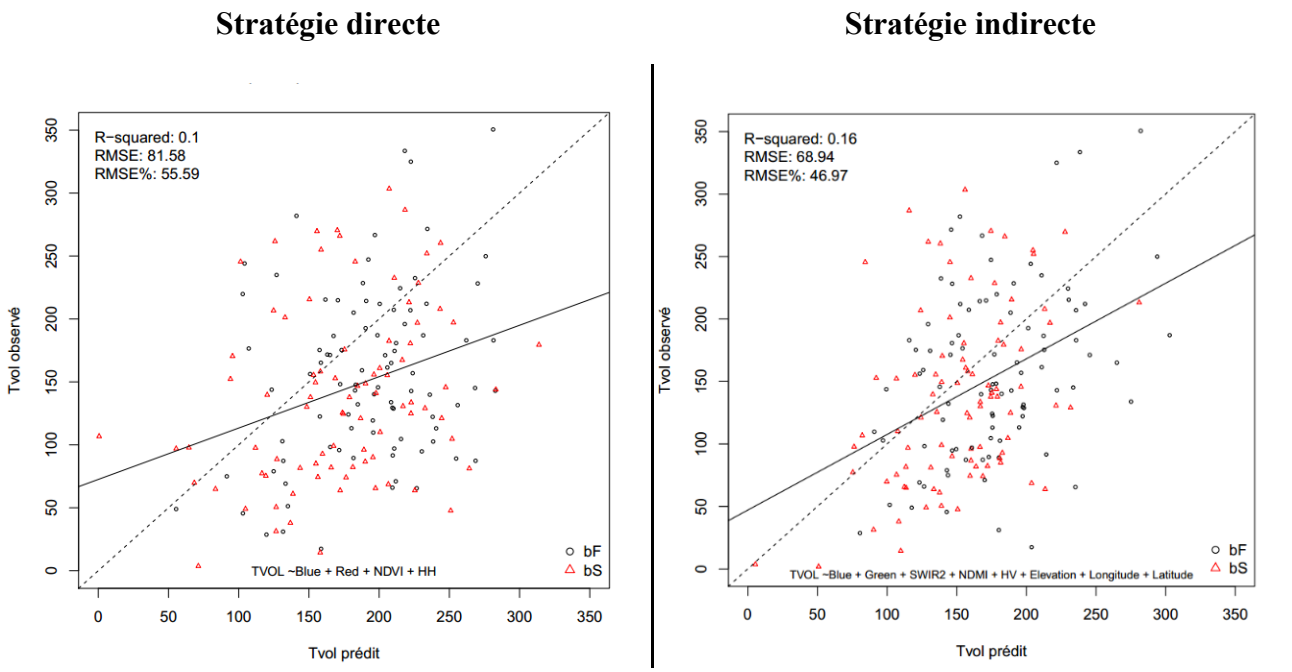


Figure 13. Ligne noire : Valeurs de Tvol observées versus prédites (m^3/ha) modélisées par la régression OLS en utilisant la stratégie directe et indirecte (phase 1 et phase 2) en utilisant les données de validation.

5.2.2 Random Forest

En ce qui concerne la modélisation à partir de l'approche RF pour la stratégie directe, le RMSEcv relatif obtenu par validation croisée lors du diagnostic du modèle est de 34,93 %, la variance de Tvol expliquée est de 19 % et le biais est quant à lui non significatif avec une valeur de -0,12 (Tableau 7). La variable prédictive présentant le plus d'influence dans le modèle RF (% incMSE) est d'abord la polarisation HH du radar PALSAR, suivi de l'indice SWIR et du NDVI (Figure 14). L'indice EVI participe aussi à l'augmentation de la pureté des nœuds avec la variable géographique de Latitude, mais légèrement moins au niveau de l'augmentation du % MSE. Les variables Longitude, Red, AP et TCBRI ont également une importance significative dans le modèle de la stratégie directe. La droite de régression entre le Tvol prédit et observé de RF a une pente similaire à 1 et l'ordonnée à l'origine ne diffère pas significativement à [0:0] (Figure 15). Cependant, les valeurs prédites présentent une distribution centralisée avec des valeurs prédites concentrées entre 125 et 250 m^3/ha approximativement (Figure 16). Lorsque le modèle est utilisé pour prédire le Tvol sur les données de validation, le R^2 diminue à 14 %, le RMSE relatif augmente à 49,48 % et le biais est alors significatif (-32,69). Le graphique des valeurs prédites versus observées présente une légère déviance de la pente et l'ordonnée à l'origine des deux pentes est similaire.

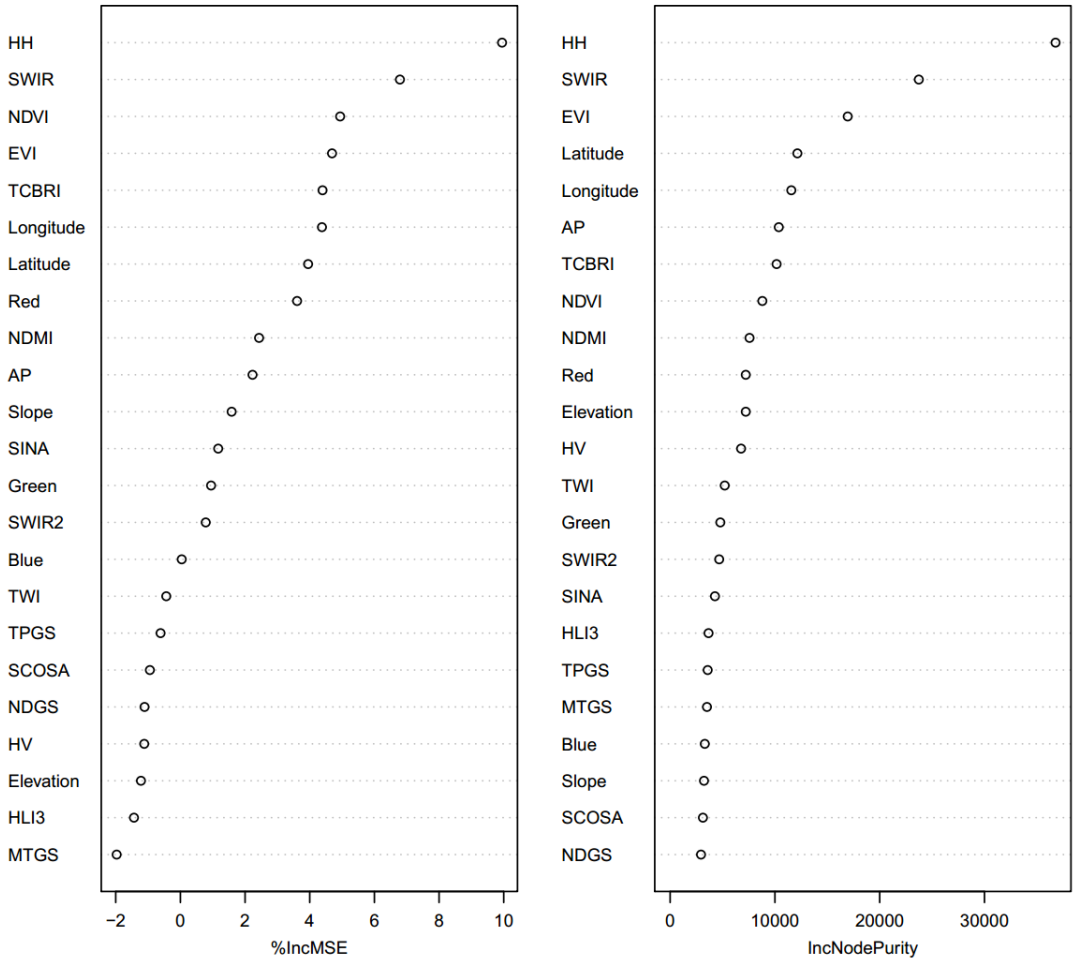


Figure 14. Importance des variables (% IncMSE et IncNodePurity) de la stratégie directe pour la modélisation RF.

Diagnostique du développement des modèles RF — Tvol (m^3/ha)

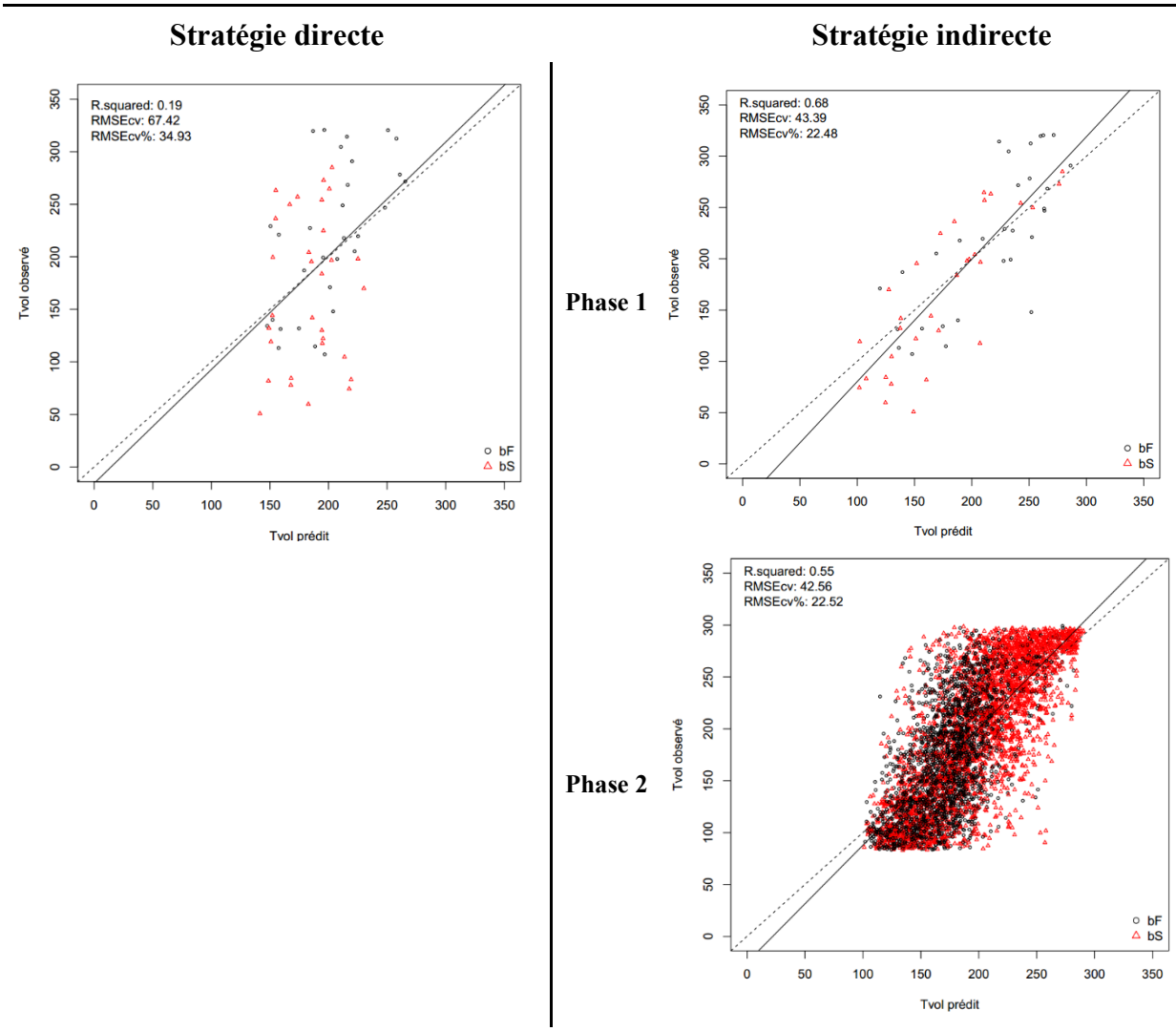


Figure 15. Ligne noire : Valeurs de Tvol observées versus prédites (m^3/ha) modélisées par la régression OLS en utilisant la stratégie directe et indirecte (phase 1 et phase 2), en fonction de la validation croisée « k-fold » pour le diagnostic du développement des modèles.

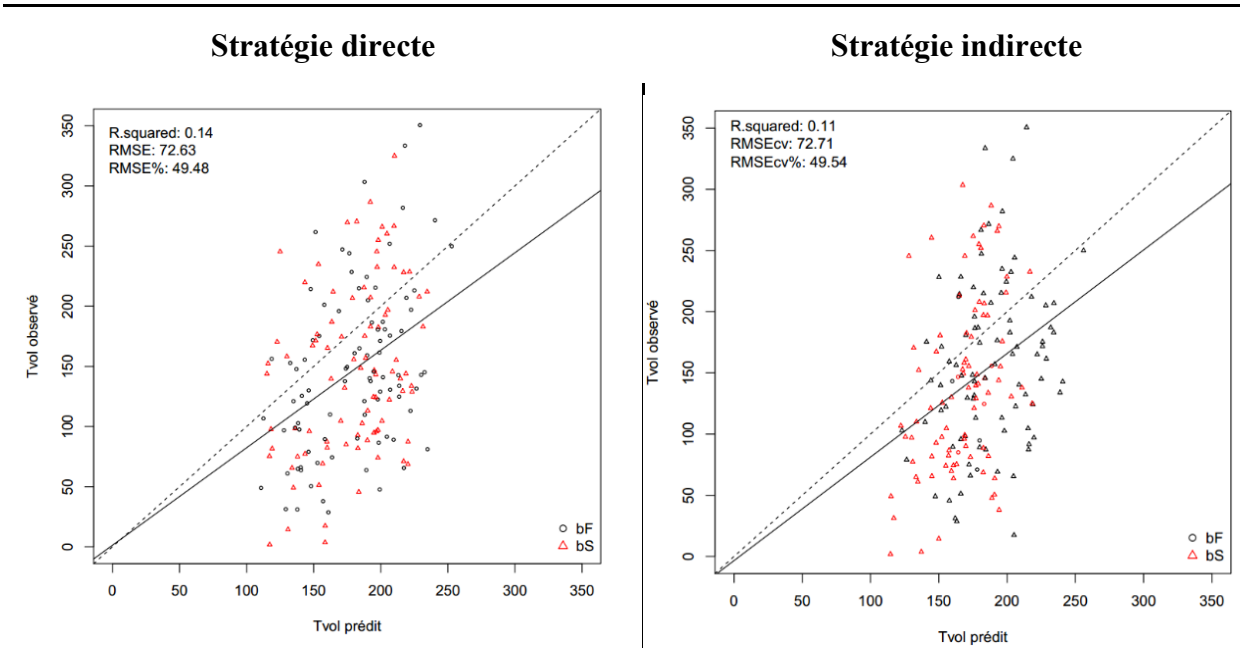


Figure 16. Ligne noire : Valeurs de Tvol observées versus prédites (m^3/ha) modélisées par la régression OLS en utilisant la stratégie directe et indirecte (phase 1 et phase 2) en utilisant les données de validation.

5.3 Stratégie indirecte

Dans l'optique d'alléger la présentation des résultats ayant trait à la stratégie de modélisation indirecte et de regrouper l'analyse des résultats des 2 phases de la stratégie indirecte, la section suivante présentera d'abord les résultats concernant le processus d'échantillonnage des placettes de substitution réalisé suite aux prédictions du Tvol dans les transects ALS par les modèles de la phase 1 de la stratégie indirecte. Les résultats de la phase 1 et de la phase 2 seront ensuite présentés pour chacune des approches de modélisation (OLS et RF).

5.3.1 Échantillonnage des placettes de substitution pour la modélisation de la phase 2 de la stratégie indirecte

Comme mentionné précédemment, la mise en place de la phase 2 de la stratégie indirecte a d'abord nécessité la sélection de variables d'équilibrage permettant d'obtenir un échantillonnage des placettes de substitution balancée dans l'espace auxiliaire pour le développement des modèles de la phase 2 de la stratégie indirecte. D'abord, les cinq variables, ayant la relation la plus forte avec la variable d'intérêt (Tvol), sélectionnée comme variable d'équilibrage sont les quatre métriques ALS suivantes :

LH25, LHMEAN RATIOMEAN, CCfhtMEAN, ainsi que la variable à couverture complète HH du radar PALSAR (Tableau 8). La corrélation des métriques ALS avec le Tvol varie de 0,74 à 0,81. La variable HH a une plus faible corrélation, soit 51 % avec le Tvol. Ensuite, une analyse de sensibilité du nombre d'échantillons d'entrainement a été effectuée pour établir le nombre de points d'échantillons requis pour des résultats stables et a déterminé qu'une asymptote de précision était atteinte à environ 5 000 placettes. La technique d'échantillonnage appliquée (lpm_kdtree) pour sélectionner les 5 000 placettes de substitution au sein des transects ALS a spatialement distribué la sélection de manière uniforme pour les 2 approches de modélisation (OLS et RF), en plus d'avoir évité de sélectionner des pixels aux bordures extérieures des transects ALS (Figure 17).

Tableau 8. Variables auxiliaires sélectionnées pour balancer l'échantillonnage avec « Local pivotal technique » pour la stratégie indirecte

Nom des métriques	**Définition**	**Corrélation avec Tvol**
HH	Polarisation HH du radar Palsar	0,51
LHMEAN	Moyenne des hauteurs de point > 2 m	0,74
LH25	25e centile de la hauteur des points > 2 m	0,80
CCfHtMEAN	Pourcentage des premiers retours au-dessus de la moyenne	0,80
RATIOMEAN	(Tous les retours au-dessus de la moyenne) / (Total des premiers retours) * 100	0,81

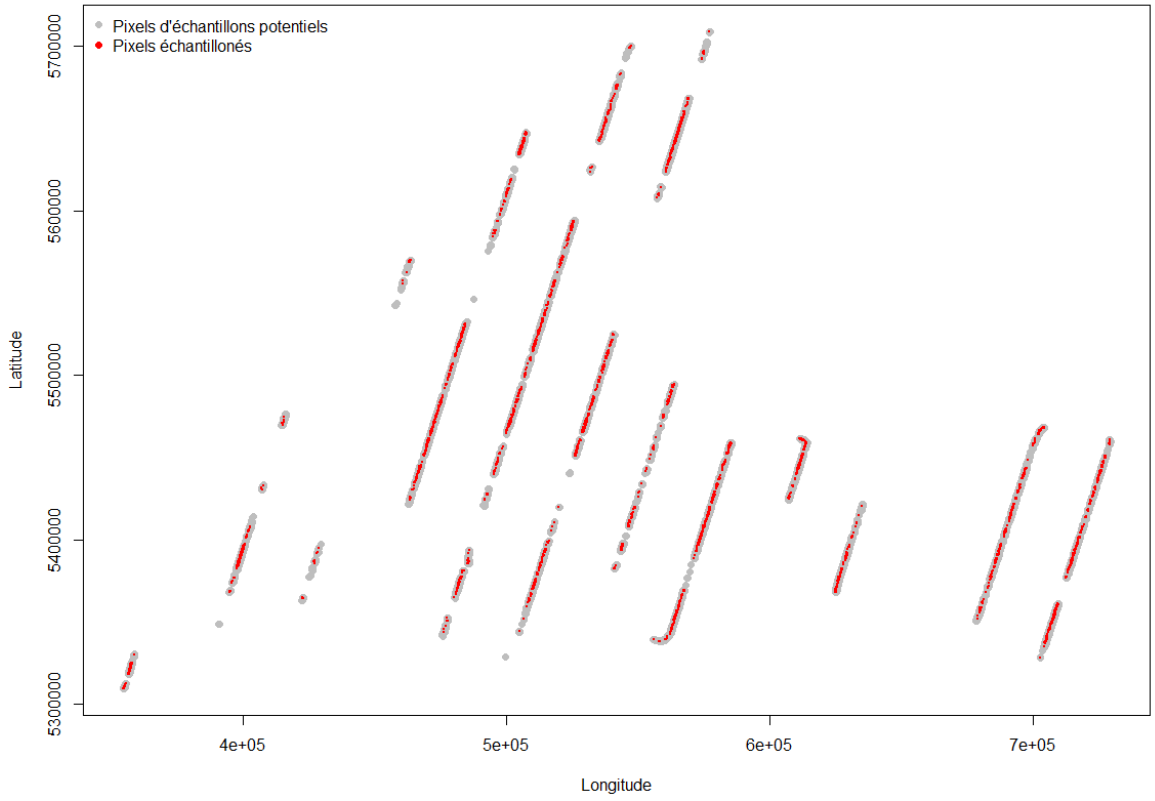


Figure 17. Exemple de la distribution spatiale des 5000 placettes de substitution échantillonnées pour la modélisation de la phase 2 de la stratégie indirecte (OLS).

L'histogramme des valeurs de placettes de substitution de Tvol échantillonnées pour la méthode de régression OLS est réparti entre 3 et 350 m^3/ha approximativement et présente une distribution presque uniforme (Figure 18 et Tableau 9) sur l'étendue de la plage des valeurs. En ce qui a trait aux valeurs de Tvol échantillonnées sur les transects ALS suite aux prédictions obtenues par la méthode RF, l'histogramme présente une faible proportion d'échantillons dans les valeurs extrêmes, dont le minimum est d'environ 80 m^3/ha et le maximum 285 m^3/ha (Figure 19). La déviation standard au sein des placettes de substitution est de 100,55 m^3/ha pour la méthode OLS, alors qu'elle est de seulement 58,77 m^3/ha pour la méthode RF (Tableau 9).

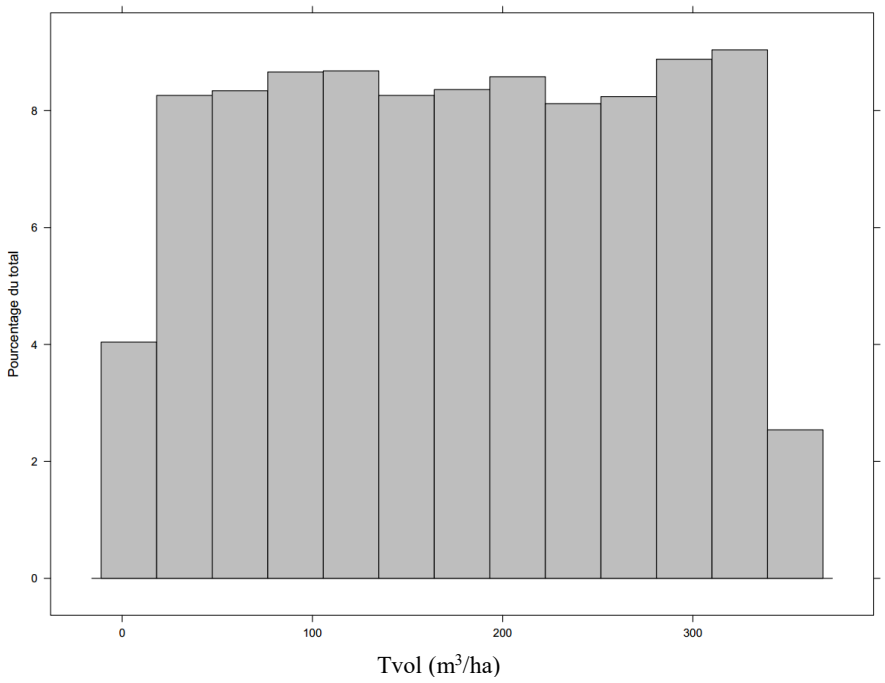


Figure 18. Histogramme des valeurs de Tvol (m3/ha) des placettes de substitution ALS de la méthode OLS.

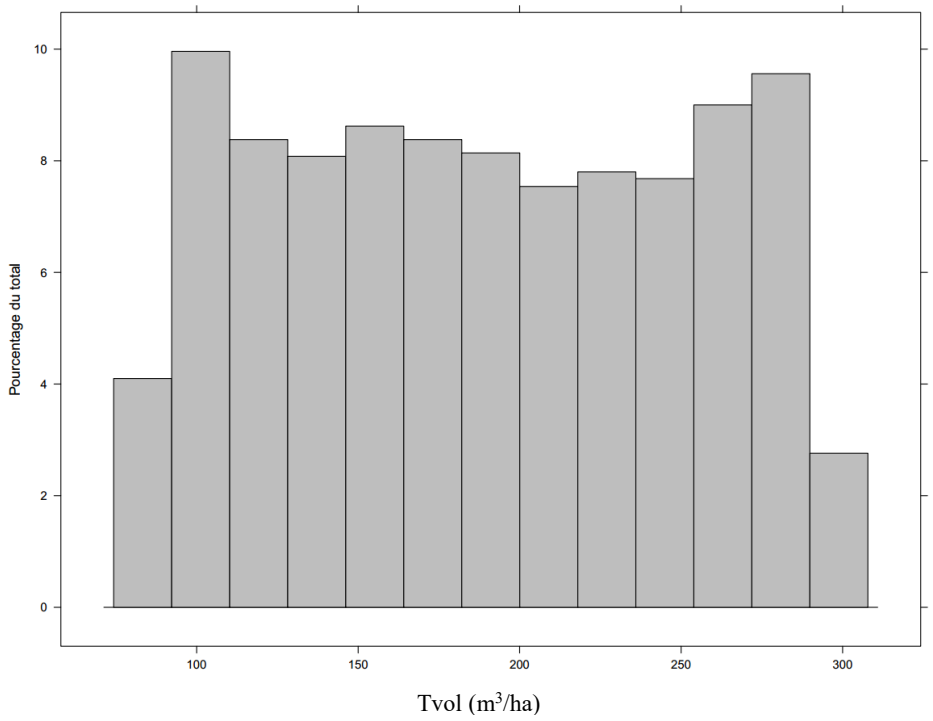


Figure 19. Histogramme des valeurs de Tvol (m^3/ha) des placettes de substitution de la méthode RF.

Tableau 9. Statistiques descriptives du volume total des placettes de substituions échantillonnées pour la méthode OLS et RF.

	Volume total (m^3/ha)			
	Min	Max	Moyenne	Écart-type
Placettes de substitution pour OLS				
Épinette noire et sapin baumier (n =5000)	3,09	354,96	178,14	100,55
Placettes de substitution pour RF				
Épinette noire et sapin baumier (n =5000)	82,51	286,08	184,35	58,77

5.3.2 Régression OLS

Le meilleur modèle de la phase 1 pour la prédiction du Tvol possède 4 métriques ALS, soit LHKURT, LHLKURT, LHCURTmeanCUBE et le CCMEAN (Tableau 5). Ce modèle explique 79 % de la variance du Tvol et l'erreur relative de prédiction est de moins de 20 % (Tableau 6). La droite de régression entre le Tvol prédit et observé de la phase 1 de la méthode OLS a une pente similaire à 1 et l'ordonnée à l'origine ne diffère pas à [0:0] (Figure 12). Le modèle a tendance à sous-estimer légèrement les valeurs fortes de Tvol pour le sapin baumier (bF) alors qu'il surestime les valeurs basses de Tvol d'épinette noire (bS). Le modèle retenu pour la phase 2 de la stratégie indirecte a un R^2 de 0,32 et son RMSEcv relatif augmente à 46,63 % (Tableau 6). Le modèle possède huit variables prédictives, dont deux géographiques, trois provenant des bandes spectrales de Landsat BAP, un indice de végétation, de topographie et la polarisation HV de PALSAR (Tableau 5). La droite de régression entre le Tvol prédit et observé de la phase 2 de la méthode OLS a une pente similaire à 1 et l'ordonnée à l'origine ne diffère pas à [0:0] (Figure 12). Cependant, le modèle présente une tendance à surestimer les basses valeurs de Tvol pour le sapin baumier (bF) et au contraire, à sous-estimer les fortes valeurs pour l'épinette noire. Les biais des modèles des deux phases de la prédiction du Tvol par la méthode OLS ne sont pas significatifs. La variance expliquée par le modèle lorsqu'il est appliqué pour prédire le Tvol sur le jeu de données de validation est de 16 %, avec un RMSE % semblable à celui obtenu lors du diagnostic du modèle par validation croisée, soit de 46,97 % (Tableau 7). Cependant, le biais de la prédiction augmente de manière significative à -18,01. Le graphique des valeurs prédites versus observées démontre un changement de pente significatif et une ordonnée à l'origine qui bascule à une valeur approximative de 50 m^3/ha (Figure 13). Les valeurs de Tvol prédites sont regroupées majoritairement entre 50 et 250 m^3/ha. Ce qui signifie que les fortes

valeurs de Tvol observées sont sous-estimées alors que le contraire est valable pour les basses valeurs de Tvol, et ce, pour les deux espèces.

5.3.3 Random Forest

En ce qui concerne la modélisation à partir de l'approche RF pour la stratégie indirecte, le RMSEcv relatif obtenu par validation croisée lors du diagnostic du modèle de la première phase est de 22,48 %, la variance de Tvol expliqué est de 68 % et le biais est quant à lui de seulement 1,41 (Tableau 6). Trois variables prédictives se démarquent quant à leur influence dans le modèle de la phase 1 : LHMIN, RATIOMEAN et LHCURTmeanCUBE (Figure 20). Les cinq variables explicatives ayant le moins d'influence sont LHKURT, RATIOMODE, LHLKURT, LHMADMODE et LHL2. Les autres métriques ont toutes une influence similaire et modérée sur la prédiction du Tvol par RF. La droite de régression entre le Tvol prédit et observé de RF a une pente légèrement supérieure à 1 et l'ordonnée à l'origine ne diffère que légèrement de [0:0] (Figure 15).

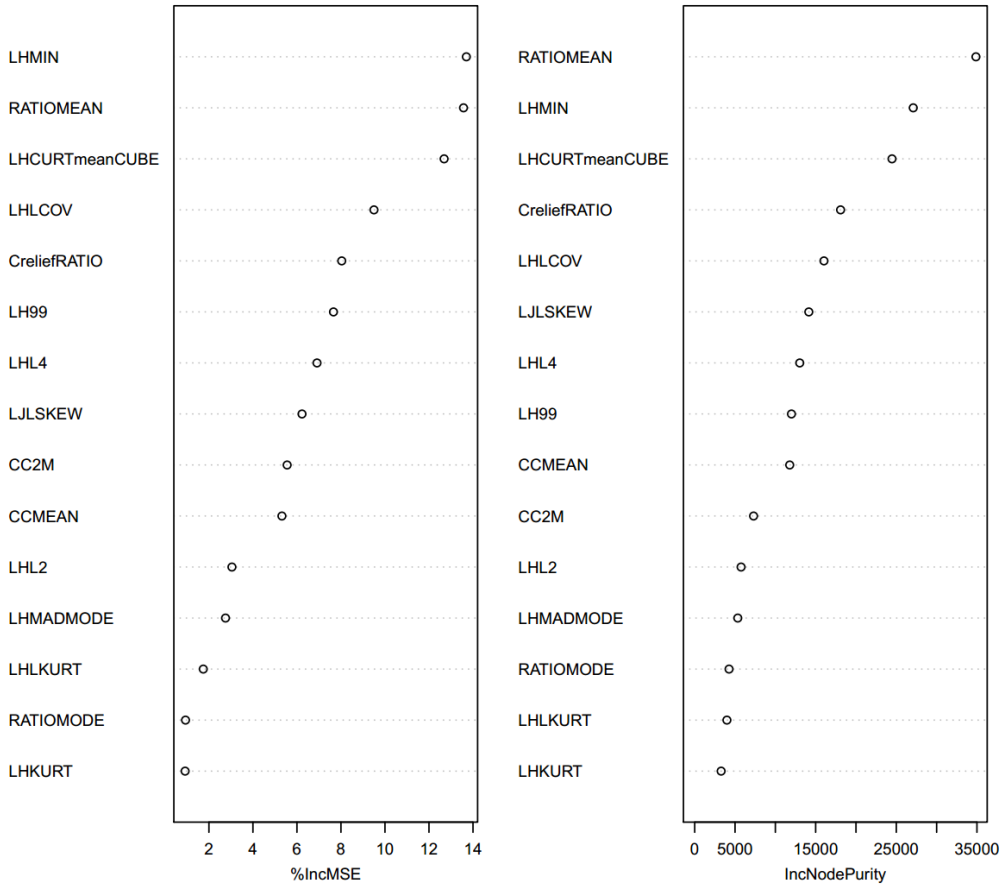


Figure 20. Importance des variables (% IncMSE et IncNodePurity) de la phase 1 de la stratégie indirecte pour la modélisation RF.

Le modèle de la phase 2 explique 55 % de la variance du Tvol et son RMSEcv relatif est de seulement 22,52 % (Tableau 6). Le biais de la phase 2 est également non significatif avec une valeur de 1,09. La variation de l'influence des variables est particulièrement importante lors de la modélisation de la phase 2 avec RF (Figure 21). Les variables ayant le plus d'importance augmentent le pourcentage de MSE entre 40 et 60 %, alors que la variable ayant le moins d'importance a une valeur de seulement 10 %. Le même phénomène s'observe quant à l'importance des variables relativement à augmentation de la pureté d'un nœud de décision. La variable Red associée à la bande spectrale rouge de Landsat BAP influence le plus la pureté d'un nœud (valeur supérieure à 1 000 000). Cette variable fait également partie de celles ayant une influence de plus de 40 % dans l'augmentation du MSE. Dans le cas des arbres de la phase 2, une dizaine de variables appartenant à quatre principales catégories (topographie, climat, indices et bandes spectrales, géographie) ont majoritairement de l'influence. La droite de régression entre le Tvol prédit et observé de RF de la phase 2 a une pente légèrement supérieure à 1 et l'ordonnée à l'origine ne diffère que légèrement de [0:0] (Figure 15). Les prédictions de Tvol obtenues par le modèle de la phase 2 sont toutes comprises entre 100 et 300 m^3/ha approximativement. Lorsque le modèle est utilisé pour prédire le Tvol sur les données de validation, le R^2 diminue à 11 %, le RMSE relatif augmente à 49,54 % et le biais est alors significatif avec une valeur de -30,72. Le graphique des valeurs prédites versus observées présente une déviance descendante de la pente, mais l'ordonnée à l'origine est similaire à [0:0] (Figure 16). Les valeurs de Tvol prédites sur les données indépendantes se situent approximativement entre 100 et 250 m^3/ha.

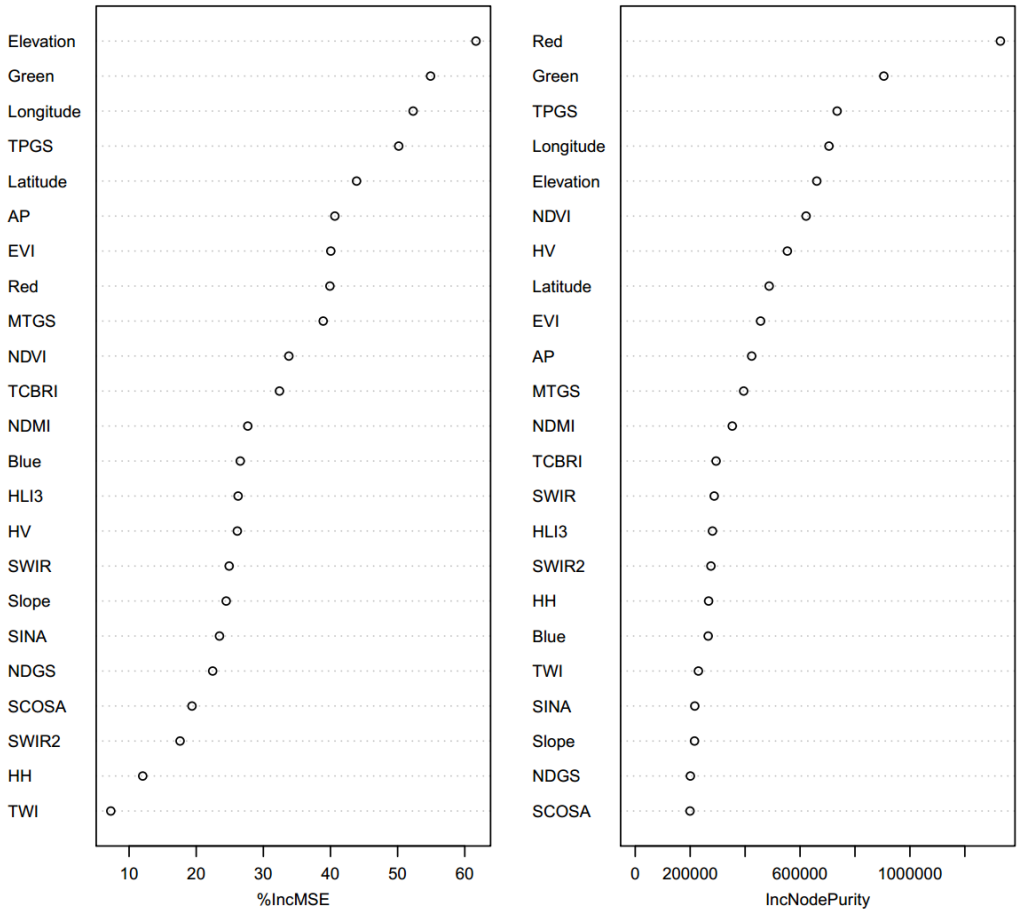


Figure 21. Importance des variables (% IncMSE et IncNodePurity) de la phase 2 de la stratégie indirecte pour la modélisation RF.

6. Discussion

6.1 Comparaison des stratégies directe et indirecte

L'objectif principal de ce projet était de développer une méthode efficace de cartographie pour l'attribut du volume total (Tvol) forestier pour la forêt boréale sur l'île de Terre-Neuve, au Canada. Le premier objectif spécifique consistait à comparer les stratégies de modélisation directe et indirecte qui combinent des placettes forestières, des transects ALS et des données à couverture complète pour cartographier l'attribut du Tvol. Les résultats obtenus ne permettent pas de confirmer la première hypothèse qui stipulait qu'une stratégie indirecte améliorerait la prédiction du Tvol par rapport à une stratégie directe. Dans notre application des méthodes, et malgré nos nombreuses tentatives d'amélioration des résultats, la stratégie indirecte, utilisant des transects de données ALS et dans laquelle deux ensembles de relations empiriques sont développés, n'a pas amélioré significativement la prédiction du Tvol par rapport à la stratégie directe qui n'utilise que des placettes terrain et des couches de données à couverture complète pour développer une seule relation empirique. La précision des résultats du Tvol des modèles de la stratégie directe est cohérente avec les résultats d'autres études et même inférieure. En effet, les erreurs relatives des modèles de validation (32-35 %) correspondent à celles obtenues par Luther et al. (2019) lors de la prédiction de l'attribut du volume total (36-39 %). Nos valeurs de RMSE relatif sont cependant inférieures à ceux mentionnés par Tomppo et al. (2008) pour les prédictions du volume effectuées au niveau du pixel par les recherches en Suisse et en Norvège (50-80 %).

La précision des résultats du Tvol de la stratégie indirecte diffère significativement de la phase 1 à la phase 2. Les modèles de phase 1 pour la prédiction de Tvol à partir de données ALS concordent avec ceux de plusieurs études similaires indiquant que les métriques dérivées des données de la ALS peuvent être utilisées pour prédire les attributs de la structure forestière (Andersen et al., 2005 ; Luther et al., 2013 ; Luther et al., 2019). Par exemple, la variance expliquée par nos modèles se situe entre 68 % et 79 % alors que la variance pour la prédiction du volume total par Luther et al. (2013) est de 66 % pour le sapin baumier et 82 % pour l'épinette noire. La nouvelle étude de Luther et al. (2019) présente de meilleurs résultats avec des R^2 supérieurs à 90 % pour la prédiction du volume total. L'amélioration des résultats pour la phase 1 de la stratégie indirecte de cette nouvelle étude peut être expliquée par l'acquisition de données ALS se chevauchant et d'une plus grande densité ainsi qu'une zone d'étude plus restreinte. L'analyse des R^2 et du RMSEcv relatif des modèles de la phase 2 de la

stratégie indirecte démontre une chute importante de précision pour l'approche OLS, mais non significative pour l'approche RF. Cela peut être dû en partie à la capacité des modèles RF à saisir les relations plus complexes entre les attributs de la forêt et les données satellitaires et environnementales, alors que les modèles de régression sont basés sur l'hypothèse de relation linéaire (Luther et al., 2019).

Les erreurs de prédiction (RMSE relatif) des modèles avec l'ensemble de données de validation étaient similaires et élevées pour les deux stratégies de prédiction du Tvol (directe : 50-56 % et indirecte : 47-50 %). Malgré des erreurs relatives élevées pour les modèles de validation des deux stratégies, la stratégie indirecte présente un RMSE relatif légèrement inférieur à celui de la stratégie directe pour l'approche OLS (47 % vs 55 % pour la stratégie directe). Ceci représente donc une amélioration légère, mais non significative, par rapport aux résultats de la stratégie directe. Les RMSE relatifs obtenus par la stratégie indirecte, pour les deux approches de modélisation (OLS et RF), sont supérieurs aux erreurs relatives de 31 % obtenues par Wikes et al. (2015) lors de la prévision de la hauteur du couvert forestier et des valeurs relatives de RMSD inférieur à 26 % des résultats de Luther et al. (2019). Dans notre étude, il y a une forte augmentation du RMSE lors de l'évaluation de la précision à partir des données de validation. En effet, le RMSEcv relatif obtenu par les statistiques de développement du modèle pour la phase 2 de la stratégie indirecte était d'environ 23 %, plutôt que de 50 % pour le modèle de RF lors de la validation. Cette valeur de RMSE relatif est inférieure à ce qu'ont obtenu Wikes et al. (2015) et Luther et al. (2019). La variabilité au sein des données de validation ou l'erreur de positionnement pourrait expliquer en partie cette forte augmentation de l'erreur. En somme, les résultats obtenus avec les données de validation ne permettent pas de confirmer la première hypothèse qui stipulait qu'une stratégie indirecte améliorerait la prédiction du Tvol par rapport à une stratégie directe.

6.2 Comparaison des approches OLS et RF

Le deuxième objectif spécifique consistait à comparer les performances des méthodes paramétriques (OLS) et non-paramétriques (RF) pour prédire et cartographier l'attribut du Tvol sur un grand territoire. Les résultats obtenus pour la prédiction du Tvol selon les approches paramétriques OLS et non-paramétrique RF n'ont pas confirmé l'hypothèse de départ selon laquelle l'approche non-paramétrique RF augmenterait la précision de la prédiction du Tvol de la forêt par rapport à la méthode OLS pour les stratégies de modélisation directe et indirecte. La performance des approches

paramétrique et non-paramétrique à prédire le Tvol dans cette étude varie en fonction de la stratégie utilisée, de la phase de la stratégie et au cours de la validation. Aucune tendance claire quant aux capacités de prédiction des deux approches ne peut donc être avancée.

En effet, en ce qui concerne la stratégie directe, les statistiques associées au diagnostic des modèles (validation croisée) suggèrent des résultats nettement différents pour les approches paramétriques et non-paramétriques pour prédire le Tvol. En effet, la variance expliquée de Tvol par l'approche OLS (44 %) est plus de deux fois supérieure à celle obtenue par l'approche RF (19 %). Cependant, le RMSEcv relatif de l'approche OLS n'est pas significativement inférieur à celui obtenu par RF. L'approche RF semble systématiquement surestimer les basses valeurs de Tvol et sous-estimer les hautes valeurs, particulièrement lorsque le modèle utilise des variables à couverture complète (comme les variables topographiques, climatiques ou les bandes spectrales), et ce peu importe la stratégie employée. Une fois que les modèles de la stratégie directe sont appliqués sur les données indépendantes de validation, la supériorité de prédiction de l'approche OLS n'est plus présente, car la différence de variance expliquée par RF et OLS est alors non significative (14 % vs 11 %).

Lors de l'implémentation de la stratégie indirecte, l'approche OLS a montré légèrement plus de variances que RF (79 % vs 68 %) lors de la phase 1 et son RMSEcv relatif est une fois de plus similaire à de celui l'approche RF (19,77 % vs 22,48 %). Cependant, l'approche RF obtient de meilleurs résultats lors de la phase 2 de la stratégie indirecte que celle de l'approche OLS. Le R^2 de OLS chute à 32 % alors que celui de RF est de 55 %. Le RMSEcv relatif de l'approche OLS double lors de la seconde phase et dépasse alors drastiquement celui obtenu par RF (46,63 % vs 22,52 %). Comme pour la stratégie directe, les résultats obtenus sur les données de validation avec la stratégie indirecte ne présentent pas de différence significative (R^2 de 16 % vs 11 % et RMSE % de 46,97 vs 49,54 %). L'incapacité de RF à prédire au-delà de la plage des données d'apprentissage pourrait expliquer la chute drastique de son R^2 et son biais élevé obtenu lors de la prédiction du Tvol à partir des données de validation. En somme, la précision des prédictions faites à partir de l'approche OLS est supérieure à celle de l'approche RF lorsqu'on regarde les statistiques des modèles de la stratégie directe et de la phase 1 de la stratégie indirecte, mais aucune différence significative d'amélioration n'est observée entre OLS et RF pour les deux stratégies lorsque les modèles sont appliqués sur les données de validation.

6.3 Limites rencontrées et sources d'erreurs

L'hypothèse de départ du projet selon laquelle la stratégie indirecte améliorerait la prédiction du Tvol par rapport à une stratégie directe n'a pu être confirmée. Plusieurs limitations et sources d'erreurs peuvent expliquer la faible performance des approches OLS et RF dans notre étude au niveau des deux stratégies. D'abord, l'étendue de la zone d'étude était nettement supérieure à celle des autres études ayant testé la stratégie indirecte (30 000 km2 vs env. 500-8000 km2) (Hudak et al., 2002 ; Andersen et al., 2012 ; Strunk et al., 2014 ; Luther et al., 2019). Le faible nombre de placettes disponibles sur l'ensemble de l'île de Terre-Neuve proportionnellement à la grandeur de la zone d'étude pour le calibrage des modèles, comparativement à celle des études précédemment citées, peut certainement avoir contribué aux erreurs des modèles. Seules 61 parcelles de terrain ont recoupé l'étendue d'acquisition des données ALS pour un territoire forestier de 30 000 km^2. Il en découle un manque de représentativité des parcelles au sol disponibles pour le développement des modèles. La grandeur de la zone d'étude de notre projet impliquait également plusieurs zones écologiques (9) possédant des caractéristiques variables. Le nombre de placettes disponibles pour calibrer les modèles pouvait varier en fonction des zones écologiques. La représentativité de certaines zones lors du calibrage des modèles était alors compromise. En comparaison, l'étude de Luther et al. (2019), qui a prédit des attributs forestiers sur une seule zone écologique, dominé par le sapin baumier, a obtenu des capacités de prédiction supérieures avec la stratégie indirecte (amélioration de la correspondance entre les valeurs observées et prédites entre 13-49 %). Elle mentionne que de prédire au-delà des conditions écologiques capturées par les données de calibration induit des inconnus et des erreurs dans les prédictions.

Ensuite, la taille variable des placettes ainsi que le manque de précision dans leur positionnement ont aussi pu participer à induire des erreurs dans les modèles de la phase 1 de la stratégie indirecte. En effet, les erreurs de positionnement des placettes terrain causent une augmentation de la variation dans la prédiction des attributs forestiers (Katila et Tomppo, 2001). Plutôt que d'utiliser des placettes rectangulaires de taille variable, l'étude de Luther et al. (2019) a utilisé des placettes circulaires positionnées avec précision, représentant une surface de 400 m^2 qui correspond à la résolution cartographique désirée à partir des données ALS. Ces conditions ont permis d'obtenir des prédictions d'attributs forestiers précises sur les zones couvertes par les données ALS lors de la phase 1 de la stratégie indirecte. Cependant, au moment du présent projet, seul le réseau des placettes permanentes acquises par le Service de gestion des forêts de Terre-Neuve et Labrador était disponible. Comme

pour le cas des placettes de calibration, l'erreur de positionnement et la variabilité au sein des placettes indépendantes peuvent avoir limité les résultats de l'étude.

Deux facteurs potentiels d'erreurs peuvent être liés aux données de télédétection, notamment aux données ALS et aux données Landsat BAP. L'étude de Luther et al. (2019), a mis en évidence que les paramètres d'acquisition des données ALS jouent un rôle important dans la précision de prédiction des modèles. Leur étude utilise des transects ALS ayant une superposition de 50 % entre la fauchée des lignes de vol, alors que notre étude utilisait seulement des lignes de vol ALS en transect spatial réparties sur l'ensemble du territoire. White et al. (2013) stipule que la superposition des transects ALS permet d'augmenter la densité d'impulsions et permet des angles de vue multiples. Ces caractéristiques augmentent la probabilité d'obtenir des retours au sol dans le couvert forestier dense et donc, d'améliorer la précision des données ALS. Un autre facteur potentiel d'erreurs implique l'imagerie satellitaire utilisée. Terre-Neuve se situe au nord du Canada et est donc plus enclin à une couverture nuageuse et à un mauvais géoréférencement. Certaines erreurs découlant du géoréférencement des données BAP et Palsar peuvent donc avoir été induites. La fréquente couverture nuageuse diminue le rendement annuel d'observation satellitaire Landsat et l'approche de composition BAP est plus difficile dans ce type d'environnement. Dans de telles circonstances, il aurait peut-être été préférable de considérer tirer parti de la série chronologique de Landsat BAP (White et al., 2014). L'étude de Luther et al. (2019) parvient à confirmer l'hypothèse que la stratégie indirecte améliore les prédictions en utilisant les données satellitaires Sentinel 2. Ces données sont acquises la même journée contrairement aux données Landsat BAP, ce qui réduit le besoin de normalisation et ainsi réduit l'erreur induite dans les modèles. L'utilisation du produit BAP de deuxième génération produit par Hermosilla et al. (2016) utilisant les segments et les points de bris (« breakpoint »), et qui fait une prédiction à partir de régression linéaire utilisant ces points afin de lisser le bruit entre les images, aurait pu contribuer à réduire les erreurs relatives causées par l'imagerie. Cependant, seul le BAP de première génération produit par White et al. (2014) était disponible au moment de l'étude. Enfin, le rééchantillonnage important des données (données de 20, 25, 30 m rééchantillonnées à 20 m) peut aussi avoir joué un rôle déterminant dans la faible précision des résultats obtenus.

Une autre source d'erreurs pouvant avoir nui à la précision de prédiction des modèles de la stratégie indirecte est liée à l'échantillonnage des placettes de substitution. En effet, bien que la technique utilisée balançait l'échantillonnage au niveau spatial ainsi que dans l'espace des variables auxiliaires,

et que la stratification forçait la représentativité de toutes les valeurs de Tvol, un effet d'autocorrélation spatiale peut tout de même avoir été induit et causé des erreurs. L'ajout d'une contrainte de distance minimale entre les échantillons pourrait améliorer les résultats de la stratégie indirecte. Enfin, l'interprétation visuelle pour l'assignation des pixels forêt / non-forêt peut avoir induit des erreurs dans les modèles de la phase 2 de la stratégie indirecte.

En somme, puisque l'hypothèse de départ est maintenant confirmée par l'étude de Luther et al. (2019), il semble clair que les enjeux, liés au faible nombre de placettes disponibles par rapport à la grandeur du territoire, l'implication de diverses écozones, la taille différente de ces placettes, le type d'imagerie satellitaire utilisé ainsi que leur géoréférencement et enfin, le rééchantillonnage important des données ont joué un rôle déterminant dans les résultats obtenus. Les produits d'imagerie satellitaire utilisés (national [BAP] et mondial [Palsar]) semblent aussi trouver leurs limites dans des utilisations plus fines et sur de plus petits territoires ou des études plus régionales comme c'est le cas pour cette étude.

7. Conclusion

Cette étude a présenté une méthode de prédiction du Tvol pour un grand territoire (30 000 km^2) selon une résolution de 20 m. L'application de cette méthode a été démontrée à l'aide de parcelles de validation indépendantes où le Tvol variait de 2 à 350 m^3/ha. Le Tvol a été prédit en utilisant deux stratégies différentes (directe et indirecte en deux phases) et deux approches de modélisation : Ordinary least squares (OLS) et Random Forest (RF). Dans la stratégie directe, des relations statistiques ont été développées entre les mesures prises dans des parcelles de terrain et des variables spatialement disponibles pour l'ensemble du territoire de Terre-Neuve à une résolution spatiale de 20 m. La stratégie indirecte comprenait deux étapes et impliquait la combinaison de mesures dans des parcelles forestières, avec les données ALS acquises en transects sur toute l'île de Terre-Neuve et des couches spatiales en format matriciel à 20 m couvrant l'île. Au cours de la phase 1 de la méthode indirecte, les relations statistiques entre les métriques ALS et les mesures des placettes terrain ont été modélisées. Dans cette phase, la prédiction des attributs de la forêt était limitée spatialement à la zone où les données ALS avaient été collectées (dans les transects). Dans la phase 2, les relations ont été modélisées entre les valeurs de Tvol échantillonnées dans les transects ALS et les données à couverture complète sur le territoire, c'est-à-dire les images de satellites optiques, les données topographiques et climatiques. Les modèles ont fourni des erreurs relatives allant de 46,97 % à 49,54 % pour la stratégie indirecte et de 49,48 % à 55,59 % pour la stratégie directe, lorsque validés avec un réseau de placettes d'inventaire forestier indépendantes, c'est-à-dire en dehors des transects ALS. Le RMSE relatif était plus bas pour le modèle de l'approche RF (49,48 %) que celui de la régression OLS (55,59 %) pour la stratégie directe, mais l'approche OLS a obtenu un RMSE relatif légèrement plus bas lors de l'utilisation de la stratégie indirecte (46,97 % vs 49,54 %). Le biais de l'approche OLS a été significativement inférieur à celui de RF pour la stratégie de modélisation indirecte (-18,01 vs -30,72). Bien que les résultats de la stratégie indirecte ne confirment pas l'hypothèse selon laquelle la stratégie indirecte améliorerait la prédiction de Tvol, les modèles présentent tout de même de légères améliorations au niveau des biais et des erreurs relatives. Il pourrait être possible d'améliorer la représentativité des échantillons sélectionnés pour la stratégie indirecte en imposant des contraintes de sélection supplémentaires, telles que l'évitement des contours et le filtrage des échantillons, afin d'inclure uniquement ceux situés dans des zones forestières relativement homogènes. En somme, cette étude soulève l'importance du géoréférencement des données, de la taille et de la forme des placettes terrain, de la densité de points des données ALS et de l'impact possible provoqué par le rééchantillonnage de l'imagerie satellitaire.

Il est donc possible que la correction des sources d'erreurs et des limites mentionnées à la section précédente puisse potentiellement suffire pour que la stratégie indirecte améliore la précision de la prédiction du Tvol.

8. Références

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory. Petrov, B.N. and Csaki, F. (eds). Akadémiai Kiado, 267–281.

Andersen, H.-E., McGaughey, R.J. and Reutebuch, S.E. (2005). Estimating forest canopy fuel parameters using LIDAR data. Remote Sensing of Environment, 94, 441–449.

Andersen, H. E., Strunk, J., Temesgen, H., Atwood, D., and Winterberger, K. (2012). Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: a case study in the boreal forests of interior Alaska. Canadian Journal of Remote Sensing, 37 (6), 596-611.

Avery, T. E., and Burkhart, H. E. (2015). Forest measurements. Waveland Press.

Baatuuwie, N. B., and Van Leeuwen, L. (2011). Evaluation of three classifiers in mapping forest stand types using medium resolution imagery: a case study in the Offinso Forest District, Ghana. African Journal of Environmental Science and Technology, 5 (1), 25-36.

Baccini, A., Friedl, M. A., Woodcock, C. E., and Warbington, R. (2004). Forest biomass estimation over regional scales using multisource data. Geophysical research letters, 31 (10).

Beaudoin, A., Bernier, P. Y., Guindon, L., Villemaire, P., Guo, X. J., Stinson, G., Bergeron, T., Magnussen, S. and Hall, R. J. (2014). Mapping attributes of Canada's forests at moderate resolution through k NN and MODIS imagery. Canadian Journal of Forest Research, 44 (5), 521-532.

Berterretche, M., Hudak, A. T., Cohen, W. B., Maiersperger, T. K., Gower, S. T., and Dungan, J. (2005). Comparison of regression and geostatistical methods for mapping Leaf Area Index (LAI) with Landsat ETM+ data over a boreal forest. Remote Sensing of Environment, 96 (1), 49-61.

Boudreau, J., Nelson, R. F., Margolis, H. A., Beaudoin, A., Guindon, L., and Kimes, D. S. (2008). Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. Remote Sensing of Environment, 112 (10), 3876-3890.

Bouvier, M., Durrieu, S., Fournier, R. A., and Renaud, J. P. (2015). Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. Remote Sensing of Environment, 156, 322-334.

Breiman, L. (2001). Random Forests. Machine Learning, 45 (1), 5-32.

Breusch, T. S., and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287-1294.

Brosofske, K. D., Froese, R. E., Falkowski, M. J., and Banskota, A. (2013). A review of methods for mapping and prediction of inventory attributes for operational forest management. Forest Science, 60 (4), 733-756.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F., and Bertini, R. (2008). Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. Remote Sensing of Environment, 112 (5), 2686–2700.

Curran, P. J. (1986). The importance of measurement error for certain procedures in remote sensing at optical wavelengths, Photogramm. Eng. Remote Sensing, 52, 229-241.

Current Results Nexus (2015). Current results research news & science facts, average weather in Newfoundland and Labrador,

http://www.currentresults.com/Weather/Canada/Newfoundland-Labrador/average-newfoundland-labrador-weather.php, Page consultée le 15 septembre 2016.

Dech, J. P., Mayhew-Hammond, S., James, A. L., & Pokharel, B. (2014). Modeling Canada yew (Taxus canadensis Marsh.) distribution and abundance in the boreal forest of northeastern Ontario, Canada. Ecological indicators, 36, 48-58.

Dodge, Y., and Valentin. Rousson. (1999). Analyse de régression appliquée. Dunod.

Evans, J. S., and Cushman, S. A. (2009). Gradient modeling of conifer species using random forests. Landscape Ecology, 24 (5), 673-683.

Falkowski, M. J., Evans, J. S., Martinuzzi, S., Gessler, P. E., and Hudak, A. T. (2009). Characterizing forest succession with LiDAR data: An evaluation for the Inland Northwest, USA. Remote Sensing of Environment, 113 (5), 946–956.

Falkowski, M. J., Hudak, A. T., Crookston, N. L., Gessler, P. E., Uebler, E. H., and Smith, A. M. (2010). Landscape-scale parameterization of a tree-level forest growth model: a k-nearest neighbor imputation approach incorporating LiDAR data. Canadian Journal of Forest Research, 40 (2), 184-199.

Freeman, E. A., Frescino, T. S., and Moisen, G. G. (2009). ModelMap: an R Package for Model Creation and Map Production. R Package Version, 4-6. URL: https://CRAN.R-project.org/package=ModelMap (Page consultée le 23 juillet 2018).

Fitzgerald, R. W., and Lees, B. G. (1993). The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain. International Archives of Photogrammetry and Remote Sensing, 29, 570-570.

Franco-Lopez, H., Ek, A. R., and Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote sensing of Environment, 77 (3), 251-274.

Frazer, G. W., Magnussen, S., Wulder, M. A., and Niemann, K. O. (2011). Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sensing of Environment, 115 (2), 636-649.

García, M., Riaño, D., Chuvieco, E., & Danson, F. M. (2010). Estimating biomass carbon stocks for a Mediterranean forest in central Spain using LiDAR height and intensity data. Remote Sensing of Environment, 114 (4), 816-830.

Gillis, M. D., Omule, A. Y., and Brierley, T. (2005). Monitoring Canada's forests: the national forest inventory. The Forestry Chronicle, 81 (2), 214-221.

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics, 24 (2), 120-131.

Hagner, O., and Reese, H. (2007). A method for calibrated maximum likelihood classification of forest types. Remote sensing of environment, 110 (4), 438-444.

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., Hobart, G. W., and Campbell, L. B. (2016). Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring. International Journal of Digital Earth, 9 (11), 1035-1054.

Holm, S., Nelson, R., and Ståhl, G. (2017). Hybrid three-phase estimators for large-area forest inventory using ground plots, airborne LiDAR, and space LiDAR. Remote sensing of environment, 197, 85-97.

Holmgren, J., Persson, Å., and Söderman, U. (2008). Species identification of individual trees by combining high resolution LiDAR data with multi-spectral images. International Journal of Remote Sensing, 29 (5), 1537-1552.

Hopkinson, C., Chasmer, L., Colville, D., Fournier, R. A., Hall, R. J., Luther, J. E., ... and St-Onge, B. (2013). Moving toward consistent ALS monitoring of forest attributes across Canada. Photogrammetric Engineering and Remote Sensing, 79 (2), 159-173.

Houghton, R. A., Butman, D., Bunn, A. G., Krankina, O. N., Schlesinger, P., and Stone, T. A. (2007). Mapping Russian forest biomass with data from satellites and forest inventories. Environmental Research Letters, 2 (4), 045-032.

Hudak, A. T., Lefsky, M. A., Cohen, W. B., and Berterretche, M. (2002). Integration of LiDAR and Landsat ETM+ data for estimating and mapping forest canopy height. Remote sensing of Environment, 82 (2), 397-416.

Hudak, A. T., Crookston, N. L., Evans, J. S., and Falkowski, M. J. (2006). Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return LiDAR and multispectral satellite data. Canadian Journal of Remote Sensing, 32, 126–138.

Hudak, A. T., Crookston, N. L., Evans, J. S., Hall, D. E., and Falkowski, M. J. (2008). Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment, 112 (5), 2232-2245. Corrigendum: Remote Sensing of Environment, (2009). 113, 1, 289-290.

Hudak, A. T., Strand, E. K., Vierling, L. A., Byrne, J. C., Eitel, J. U., Martinuzzi, S., & Falkowski, M. J. (2012). Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. Remote Sensing of Environment, 123, 25-40.

Hyyppä, J., Yu, X., Hyyppä, H., Vastaranta, M., Holopainen, M., Kukko, A., ... & Alho, P. (2012). Advances in forest inventory using airborne laser scanning. Remote sensing, 4 (5), 1190-1207.

Ingram, J. C., Dawson, T. P., and Whittaker, R. J. (2005). Mapping tropical forest structure in southeastern Madagascar using remote sensing and artificial neural networks. Remote Sensing of Environment, 94 (4), 491-507.

Jollife, I.T. 2002. Principal Component Analysis. 2nd ed. New York: Springer-Verlag.

Kaiser, Henry F. 1961. “A Note on Guttman’s Lower Bound for the Number of Common Factors.” British Journal of Statistical Psychology 14: 1–2.

Katila, M., and Tomppo, E. (2001). Selecting estimation parameters for the Finnish multisource National Forest Inventory. Remote Sensing of Environment, 76 (1), 16-32.

Ker, M.F. (1974). Metric Yield Tables for the Major Forest Cover Types of Newfoundland. Inf. Rep. M-X-141. Natural Resources Canada, Canadian Forest Service–Atlantic Forestry Centre.

Labrecque, S., Fournier, R. A., Luther, J. E., and Piercey, D. (2006). A comparison of four methods to map biomass from Landsat-TM and inventory data in western Newfoundland. Forest Ecology and Management, 226 (1-3), 129-144.

Leboeuf, A. and R.A. Fournier (2015). Mapping vegetation and surficial deposits of the northern boreal forests in Québec using satellite images to report its status and resilience to forest fire. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS).8 (11), 5199-5211.

Lefsky, M. A., Hudak, A. T., Cohen, W. B., & Acker, S. A. (2005). Geographic variability in LiDAR predictions of forest stand structure in the Pacific Northwest. Remote Sensing of Environment, 95 (4), 532-548.

Le Maire, G., Marsden, C., Nouvellon, Y., Grinand, C., Hakamada, R., Stape, J. L., and Laclau, J. P. (2011). MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. Remote Sensing of Environment, 115 (10), 2613-2625.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R news, 2 (3), 18-22.

Liaw, A., and Wiener, M. (2002). Package "randomForest", R package version 4.6-14. URL: https://CRAN.R-project.org/package=leaps (Page consultée le 29 juillet 2017).

Lim, T. K. (2006). A Landsat surface reflectance dataset for North America, 1990-2000. IEEE Geoscience and Remote Sensing Letters, 3 (1), 68-72.

Lisic, J et Grafström A. (2018). Package "SamplingBigData". R package Version 1. URL: https://CRAN.R-project.org/package=SamplingBigData (Page consultée le 25 juillet 2018).

Lumley, T. (2017). Package « leaps » : regression subset selection. R package version 3.0. URL: http://CRAN.R-project.org/package=leaps. (Page consultée le 15 mars 2017).

Luther, J. E., Fournier, R. A., Piercey, D. E., Guindon, L., and Hall, R. J. (2006). Biomass mapping using forest type and structure derived from Landsat TM imagery. International journal of applied earth observation and geoinformation, 8 (3), 173-187.

Luther, J. E., Skinner, R., Fournier, R. A., van Lier, O. R., Bowers, W. W., Coté, J. F., ... and Moulton, T. (2013). Predicting wood quantity and quality attributes of balsam fir and black spruce using airborne laser scanner data. Forestry, 87 (2), 313-326.

Luther, J. E., Fournier, R. A., van Lier, O. R., and Bujold, M. (2019). Extending ALS-Based Mapping of Forest Attributes with Medium Resolution Satellite and Environmental Data. Remote Sensing, 11 (9), 1092.

Mahoney, C., Hall, R., Hopkinson, C., Filiatrault, M., Beaudoin, A., and Chen, Q. (2018). A forest attribute mapping framework: a pilot study in a northern boreal forest, Northwest Territories, Canada. Remote Sensing, 10 (9), 1338.

Mallows, C. L. (1973). Some comments on C p. Technometrics, 15 (4), 661-675.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., and Kangas, J. (2006). Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. Canadian Journal of Forest Research, 36 (2), 426-436.

Masek, J. G., Vermote, E. F., Saleous, N. E., Wolfe, R., Hall, F. G., Huemmrich, K. F., ... and Lim, T. K. (2006). A Landsat surface reflectance dataset for North America, 1990-2000. IEEE Geoscience and Remote Sensing Letters, 3 (1), 68-72.

Maselli, F., Chiesi, M., Montaghi, A., and Pranzini, E. (2011). Use of ETM+ images to extend stem volume estimates obtained from LiDAR data. ISPRS journal of photogrammetry and remote sensing, 66 (5), 662-671.

Matasci, G., Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., Hobart, G. W., and Zald, H. S. (2018). Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and LiDAR plots. Remote sensing of environment, 209, 90-106.

Mazerolle M.J. (2006) Improving data analysis in herpetology: using Akaike' s information criterion (AIC) to assess the strength of biological hypotheses. Amphibia-Reptilia. 27: 169-180.

McCune, B., and Keon, D. (2002). Equations for Potential Annual Direct Incident Radiation and Heat Load. Journal of Vegetation Science, 13 (4), 603–606.

McGaughey, R. (2014). FUSION/LDV: Software for LiDAR data analysis and visualization. Version 3.41. Seattle, WA: U.S. Department of Agriculture, Forest Service, Pacific NorthwestResearch Station.

McInerney, D. O., Suarez-Minguez, J., Valbuena, R., and Nieuwenhuis, M. (2010). Forest canopy height retrieval using LiDAR data, medium-resolution satellite imagery and kNN estimation in Aberfoyle, Scotland. Forestry, 83 (2), 195-206.

McKenney, D., Papadopol, P., Lawrence, K., Campbell, K., and Hutchinson, M. (2007). Customized spatial climate models for Canada. Frontline Note, (108).

McRoberts, R. E., Wendt, D. G., Nelson, M. D. and Hansen, M. D. (2002). Using a land cover classifi cation based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sensing of Environment, 81, 36-44.

McRoberts, R. E. (2006). A model-based approach to estimating forest area. Remote Sensing of Environment, 103 (1), 56-66.

McRoberts, R. E. (2010). Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sensing of Environment, 114 (5), 1017-1025.

McRoberts, R.E. (2011) Satellite image-based maps: Scientific inference or pretty pictures? Remote Sens. Environ. 115, 715–724.

Means, J. E., Acker, S. A., Fitt, B. J., Renslow, M., Emerson, L., & Hendrix, C. J. (2000). Predicting forest stand characteristics with airborne scanning LiDAR. Photogrammetric Engineering and Remote Sensing, 66 (11), 1367-1372.

Milne, T., Monette, S., Luther, J. E., Fournier, R., Hopkinson, C., and Bowers, W. (2012). LiDAR processing software in support of the Newfoundland fibre inventory project. In 33rd Canadian Symposium on Remote Sensing, 11–14 June 2012.

Murphy, M. A., Evans, J. S., & Storfer, A. (2010). Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. Ecology, 91 (1), 252-261.

Myers, R.H. (1990). Classical and Modern Regression with Applications, 2nd edn. PWS-KENT Publishing Company.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sensing of Environment, 80, 88-99.

Næsset, E. (2004). Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research, 19 (2), 164-179.

Næsset, E. (2007). Airborne laser scanning as a method in operational forest inventory: status of accuracy assessments accomplished in Scandinavia. Scandinavian Journal of Forest Research, 22 (5), 433-442.

Newfoundland Forest Service (2012). Geographic Information System Forest Inventory Database. Government of Newfoundland and Labrador, Department of Natural Resources. Newfoundland and Labrador.

Newfoundland government (2015). Hydrology and climate of Newfoundland. In Department of environment and conservation, URL: http://www.env.gov.nl.ca/env/waterres/cycle/hydrologic/nl.html.

Nilsson, M. (1997). Estimation of forest variables using satellite image data and airborne LiDAR. SLU.

Niska, H., Skon, J. P., Packalen, P., Tokola, T., Maltamo, M., and Kolehmainen, M. (2010). Neural networks for the prediction of species-specific plot volumes using airborne laser scanning and aerial photographs. IEEE Transactions on Geoscience and Remote Sensing, 48 (3), 1076-1085.

Ohmann, J. L., and Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Canadian Journal of Forest Research, 32 (4), 725-741.

Packalén, P., and Maltamo, M. (2006). Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. Forest Science, 52 (6), 611-622.

Pascual, C., Garcia-Abril, A., Cohen, W. B., and Martin-Fernandez, S. (2010). Relationship between LiDAR-derived forest canopy height and Landsat images. International Journal of Remote Sensing, 31 (5), 1261-1280.

Penner, M., Pitt, D. G., and Woods, M. E. (2013). Parametric vs. nonparametric LiDAR models for operational forest inventory in boreal Ontario. Canadian Journal of Remote Sensing, 39 (5), 426-443.

Peres-Neto, Pedro R., Donald A. Jackson, and Keith M. Somers. 2005. “How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited.” British Journal of Statistical Psychology 49: 974–97.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol.53 (5), 793-808.

Powell, S. L., Cohen, W. B., Healey, S. P., Kennedy, R. E., Moisen, G. G., Pierce, K. B., and Ohmann, J. L. (2010). Quantification of live aboveground forest biomass dynamics with Landsat time-

series and field inventory data: A comparison of empirical modeling approaches. Remote Sensing of Environment, 114 (5), 1053-1068.

Prasad, A. M., Iverson, L. R., Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems, 9 (2), 181-199.

R. van Lier, O., and Luther, J. 2017). Newfoundland Data for Evaluating LiDAR as a Sampling Tool for Large-Area Forest Characterization, Data Dictionary.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., and Roberts, D. (2008). Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. Remote Sensing of Environment, 112 (5), 2272-2283.

Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 115-124.

Rowe, J. S., and Halliday, W. E. D. (1972). Forest regions of Canada.

Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., ... and Hyyppä, J. (2015). Model-assisted estimation of growing stock volume using different combinations of LiDAR and Landsat data as auxiliary information. Remote Sensing of Environment, 158, 431-440.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6 (2), 461-464.

Silva, C. A., Klauberg, C., e Carvalho, S. D. P. C., & Hudak, A. T. (2014). Mapping aboveground carbon stocks using LiDAR data in Eucalyptus spp. plantations in the state of São Paulo, Brazil. Scientia Forestalis. 42 (104): 591-604., 42 (104), 591-604.

Strunk, J. L., Temesgen, H., Andersen, H.-E., and Packalen, P. (2014). Prediction of Forest Attributes with Field Plots, Landsat, and a Sample of LiDAR Strips. Photogrammetric Engineering and Remote Sensing, 80 (2), 143–150.

Stumpf, A., and Kerle, N. (2011). Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment, 115 (10), 2564-2577.

Symonds M.R.E. and Moussalli A. (2011) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike' s information criterion. Behavioral Ecology and Sociobiology. 65: 13-21.

Thenkabail, P. S., Enclona, E. A., Ashton, M. S., Legg, C., and De Dieu, M. J. (2004). Hyperion, IKONOS, ALI, and ETM+ sensors in the study of African rainforests. Remote Sensing of Environment, 90 (1), 23-43.

Tomppo, E. (1991). Satellite imagery-based national inventory of Finland. Intdernational archives of Photogrammetry, Remote Sensing. 28, 419–424.

Tomppo, E., Goulding, C., and Katila, M. (1999). Adapting Finnish multi-source forest inventory techniques to the New Zealand preharvest inventory. Scandinavian Journal of Forest Research,14, 182-192.

Tomppo, E., Korhonen, K. T., Heikkinen, J., and Yli-Kojola, H.(2001). Multisource inventory of the forests of the Hebei Forestry Bureau, Heilongjiang, China. Silva Fennica, 35, 309 - 328.

Tomppo, E., and Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. Remote Sensing of Environment, 92 (1), 1-20.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., and Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sensing of Environment, 112 (5), 1982-1999.

Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R. E., Gabler, K., Schadauer, K., ... and Cienciala, E. (2010). National forest inventories. Pathways for Common Reporting. European Science Foundation, 541-553.

Tonolli, S., Dalponte, M., Neteler, M., Rodeghiero, M., Vescovo, L., and Gianelle, D. (2011). Fusion of airborne LiDAR and satellite multispectral data for the estimation of timber volume in the Southern Alps. Remote Sensing of Environment, 115 (10), 2486-2498.

Tuominen, S., Fish, S., and Poso, S. (2003). Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory. Canadian Journal of Forest Research, 33 (4), 624-634.

U.S. Geological Survey. (2017) PRODUCT GUIDE. LANDSAT SURFACE REFLECTANCE-DERIVED. SPECTRAL INDICES. Version 3.5.

Van Aardt, J. A., Wynne, R. H., and Scrivani, J. A. (2008). LiDAR-based mapping of forest volume and biomass by taxonomic group using structurally homogenous segments. Photogrammetric Engineering & Remote Sensing, 74 (8), 1033-1044.

Warren, G.R., and Meades, J.P. (1986). Wood defect and density studies of living trees. II. Total and net volume equations for Newfoundland forest management units. Information Report N — X-242. Natural Resources Canada, Canadian Forestry Service–Newfoundland Forestry Centre.

White, J. C., Wulder, M. A., Varhola, A., Vastaranta, M., Coops, N. C., Cook, B. D., ... and Woods, M. (2013). A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. The Forestry Chronicle, 89 (6), 722-723.

White, J. C., Wulder, M. A., Hobart, G. W., Luther, J. E., Hermosilla, T., Griffiths, P., ... and Guindon, L. (2014). Pixel-based image compositing for large-area dense time series applications and science. Canadian Journal of Remote Sensing, 40 (3), 192-212.

Wilkes, P., Jones, S., Suarez, L., Mellor, A., Woodgate, W., Soto-Berelov, M., ... & Skidmore, A. (2015). Mapping forest canopy height across large areas by upscaling ALS estimates with freely available satellite data. Remote sensing, 7 (9), 12563-12587.

Woods, M., Pitt, D., Penner, M., Lim, K., Nesbitt, D., Etheridge, D., and Treitz, P. (2011). Operational implementation of a LiDAR inventory in Boreal Ontario. The Forestry Chronicle, 87 (4), 512-528.

Wulder, M.A., Seemann, D. (2003) Forest inventory height update through the integration of LiDAR data with segmented Landsat imagery. Canadian Journal of Remote Sensing 29, 536–543.

Wulder, M. A., White, J. C., Cranny, M., Hall, R. J., Luther, J. E., Beaudoin, A., and Dechka, J. A. (2008). Monitoring Canada's forests. Part 1: Completion of the EOSD land cover project. Canadian Journal of Remote Sensing, 34, 549.

Wulder, M. A., White, J. C., Nelson, R. F., Næsset, E., Ørka, H. O., Coops, N. C., ... and Gobakken, T. (2012). LiDAR sampling for large-area forest characterization: A review. Remote Sensing of Environment, 121, 196-209.

Wulder, M. A., White, J. C., Bater, C. W., Coops, N. C., Hopkinson, C., and Chen, G. (2012). LiDAR plots—A new large-area data collection option: Context, concepts, and case study. Canadian Journal of Remote Sensing, 38 (5), 600-618.

Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., ... and Dwyer, J. (2019). Current status of Landsat program, science, and applications. Remote sensing of environment, 225, 127-147.

Zald, H. S., Wulder, M. A., White, J. C., Hilker, T., Hermosilla, T., Hobart, G. W., and Coops, N. C. (2016). Integrating Landsat pixel composites and change metrics with LiDAR plots to predictively map forest structure and aboveground biomass in Saskatchewan, Canada. Remote Sensing of Environment, 176, 188-201.